Multi- and Megavariate Data Analysis

Basic Principles and Applications

Third revised edition

By MKS Umetrics AB

1992-2013 MKS Umetrics AB, all rights reserved

Information in this document is subject to change without notice and does not represent a commitment on part of MKS Umetrics. The software, which includes information contained in any databases, described in this document is furnished under license agreement or nondisclosure agreement and may be used or copied only in accordance with the terms of the agreement. It is against the law to copy the software except as specifically allowed in the license or nondisclosure agreement. No part of this user guide may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, for any purpose, without the express written permission of MKS Umetrics.

Umetrics patents and trademarks:

SEQUENTIAL MODELING, US-7523384

MONITORING MULTIVARIATE PROCESSES, US-5949678

TOP-DOWN CLUSTERING/PLS-TREE®, US-8244498

DESIGN SPACE, US-8086327

MPMC, Pending

RED-MUP®, VALUE FROM DATA®, OPLS®, O2PLS®, OPLS-DA®, O2PLS-DA®

PLS-TREE®, S-PLOT®, EZinfo®, FABSTAT®, SIMCA®. MODDE®

ID #1082

User guide edition date: May 13, 2013

MKS Umetrics AB
Stortorget 21
SE-211 34 Malmö
Sweden
Phone: +46 (0)40 664 2580
Email: info@umetrics.com

Contents

Contents i

0 Preface 1

0.1 Why MultiVariate Data Analysis (MVDA) ? 1
0.2 Projection Methods (Principal Component Analysis, PCA, and Projections to Latent Structures, PLS) 1
0.3 Recognition of differences between multi- and megavariate analysis 1
0.4 Problems areas where MVDA is useful 2
0.5 Objectives of the book 3
0.6 Organization of the book 3
0.6.1 Section I, Chapters 1 – 6 4
0.6.2 Section II, Chapters 7 – 11 4
0.6.3 Section III, Chapters 12 – 14 4
0.6.4 Section IV, Chapters 15 – 18 4
0.6.5 Section V, Chapters 19 – 22 4

1 Introduction 5

1.1 Objective 5
1.2 Introduction 5
1.3 Pattern recognition 6
1.4 Basic types of data analytical questions 6
1.4.1 Data overview 6
1.4.2 Classification and discrimination 7
1.4.3 Regression modeling 7
1.5 Illustrations of problem types – Overview of a data set 7
1.5.1 A getting started data set (FOODS) 7
1.5.2 A quality control data set (CUPRUM) 9
1.6 Illustration of problem types – Classification & Discrimination 10
1.6.1 A classical dataset in statistics (IRIS) 10
1.6.2 A second dataset for classification and discrimination (ARCHAE) 12
1.7 Illustration of problem types – Regression modeling 13
1.7.1 A multiple response optimization dataset (LOWARP) 13
1.7.2 A multivariate calibration dataset (SUGAR) 14
1.8 Questions for Chapter 1 16
1.9 Summary and discussion 16

2 Basic concepts and principles of projections 17

2.1 Objective 17
2.2 Data 17
2.2.1 A table of data is not the same as information 17
2.2.2 Variability and types of data 17
2.3 Some properties and challenges of multivariate data 19

2.3.1 All data are needed 19
2.3.2 Other features 19
2.4 Confusion of correlation and causation 20
2.5 How do we analyze our data today? 21
2.5.1 The COST approach 21
2.5.2 Classical methods of statistics 22
2.5.3 Multivariate data analysis 22
2.6 The principles of projections 23
2.6.1 K-space and M-space 23
2.6.2 Example K = 2 23
2.6.3 Example K = 3 24
2.6.4 Each row of a data table is interpretable as a point in a multidimensional space 25
2.6.5 A data table forms a point-swarm 25
2.6.6 Constructing a window into the three-dimensional space 26
2.6.7 Projections in higher-dimensional space 26
2.6.8 Score plots and loading plots 27
2.6.9 Summary of principles of projections 28
2.7 The model concept and theoretical basis for model building 28
2.7.1 The model concept 28
2.7.2 Empirical, semi-empirical, and theoretical models 29
2.7.3 Semi-empirical modeling – Taylor series expansions 29
2.7.4 Conceptual basis of semi-empirical modeling 30
2.7.5 Basic conceptual model of projections 30
2.7.6 Summary of modeling 31
2.8 Questions for Chapter 2 32
2.9 Summary and discussion 32

3 PCA 33

3.1 Objective 33
3.2 Introduction to PCA 33
3.3 Pre-treatment of data 34
3.3.1 Scaling 34
3.3.2 Mean-centering 37
3.4 A geometric interpretation of PCA 37
3.4.1 Setting up K-dimensional space 37
3.4.2 Plotting the observations in K-dimensional space 38
3.4.3 The effect of mean-centering 38
3.4.4 The first principal component 39
3.4.5 Extending the model with the second principal component 40
3.4.6 Two principal components define a model plane 41
3.4.7 How to interpret the score plot 42
3.4.8 Extensions to higher-order components 43
3.4.9 Summary of PCA 45
3.5 Additional PCA diagnostics 46
3.5.1 Observation diagnostics – Are there outliers in the data? 46
3.5.2 Variable diagnostics – Which variables are well-explained? 49
3.5.3 Model diagnostics – How many principal components are really needed? 50
3.5.4 Cross-validation 52
3.6 Questions for Chapter 3 53
3.7 Summary and discussion 53

4 PLS **55**

4.1 Objective 55
4.2 Introduction to PLS....55
4.2.1 Some terminology 55
4.2.2 Common application areas 56
4.2.3 A historical note....56
4.2.4 Theoretical foundation....57
4.3 Pre-processing of data 57
4.4 The geometry of PLS in the case of one response (M = 1) 57
4.4.1 Constructing the X and Y co-ordinate systems....58
4.4.2 The first PLS component (M = 1)....58
4.4.3 Extending the model with the second component (M = 1)....60
4.4.4 Example: Sugar....61
4.5 The geometry of PLS with multiple responses (M > 1)....63
4.5.1 Setting up the X- and Y-spaces 63
4.5.2 Calculating the first PLS component (M > 1) 64
4.5.3 Adding the second component (M > 1)....65
4.5.4 Which original variables contribute to the formation of the model planes?....66
4.5.5 Higher-order components (M > 1)....66
4.5.6 Example: LOWARP....67
4.5.7 Summary of PLS projections, scores and weights....69
4.6 PLS model interpretation....70
4.6.1 PLS weights 70
4.6.2 PLS regression coefficients....72
4.6.3 The variable influence on projection, VIP, parameter....74
4.7 Additional PLS diagnostics....76
4.7.1 Observation diagnostics – Are there non-linearities and/or outliers in the data? 76
4.7.2 Variable diagnostics – Which predictors are important and which responses are well explained?....80
4.7.3 Model diagnostics – How many PLS components should be used? ...82
4.7.4 Standard Errors and Confidence Intervals....85
4.8 Use of PLS model: Predictions 85
4.9 Questions for Chapter 4 86
4.10 Summary and discussion 86
4.10.1 Assumptions of PLS....86
4.10.2 One or several response variables? 86
4.10.3 Diagnostic tools....87
4.10.4 Many variables and many observations....87
4.10.5 Related methods 87
4.10.6 Extensions of PLS....88

5 Orthogonal PLS (OPLS) **89**

5.1 Objective 89
5.2 Introduction....89
5.3 An example with predictive and orthogonal variations....91
5.3.1 The BeerNIR dataset 91
5.3.2 An overview PCA model 91
5.3.3 PLS and OPLS models for alcohol content....93
5.4 Theory of OPLS....96
5.4.1 Objective of OPLS....96
5.4.2 The single-Y OPLS model....96
5.4.3 The multi-Y OPLS model 97
5.4.4 Y-related profiles 97
5.4.5 Implementation in SIMCA 97

5.5 A multi-Y calibration study 98
5.5.1 The Metal-Ion Complexes dataset 99
5.5.2 OPLS and PLS multi-Y models 99
5.5.3 Predictions for the test set 100
5.5.4 Interpretation of the single orthogonal component of the 4+1 OPLS model 101
5.5.5 Interpretation of the predictive components of the 4+1 OPLS model using loadings 101
5.5.6 Interpretation of the 4+1 OPLS model using coefficients 102
5.5.7 Interpretation of the predictive part of the 4+1 OPLS model using coefficients rotated into pure spectrum profiles 104
5.5.8 Interpretation of the 4+1 OPLS model using VIP 106
5.5.9 Summary; model interpretation in the case of low degree of orthogonal variation 107
5.6 What happens when the orthogonal variation increases? 107
5.6.1 Exclusion of the Ni^{2+}-response from Y (three Y-variables) 107
5.6.2 Interpretation of the first orthogonal component of the 3+2 OPLS model 108
5.6.3 Interpretation of the predictive components of the 3+2 OPLS model using loadings 108
5.6.4 Interpretation of the predictive part of the 3+2 OPLS model using coefficients rotated into pure spectrum profiles 110
5.6.5 Interpretation of the 3+2 OPLS model using VIP 111
5.6.4 Discussion of example 112
5.7 Questions for Chapter 5 112
5.8 Summary and discussion 112
5.8.1 Some general remarks 112
5.8.2 The great asset of OPLS: Improved model transparency and interpretability 113
5.8.3 Specific uses of OPLS 114
5.8.4 When and when not to use OPLS 114
5.8.5 Related methods and some extensions 114
5.8.6. Concluding remarks 115

6 O2PLS 117

6.1 Objective 117
6.2 Introduction 117
6.3 Theory of O2PLS 118
6.3.1 Objective and structure of O2PLS 118
6.3.2 The O2PLS model 119
6.3.3 Implementation in SIMCA 120
6.4 Integrating spectral and process data of a batch process 121
6.4.1 The batch hydrogenation data set 121
6.4.2 O2PLS analysis of the spectral and process data 122
6.4.3 Interpretation of the joint variation 122
6.4.4 Interpretation of the unique variation in the spectral data 124
6.4.5 Interpretation of the unique variation in the process data 125
6.4.6 Summary of the batch hydrogenation data set 126
6.5 Integrating multiple spectral datasets in a multivariate calibration study 127
6.5.1 Analyzing the joint variation in the spectral data 128
6.5.2 Analyzing the unique variation in the spectral data 128
6.5.3 Summary of the multiple spectral datasets example 129
6.6 Questions for Chapter 6 130
6.7 Summary and discussion 130

7 Multivariate characterization 131

7.1 Objective 131
7.2 Introduction....131
7.3 Rationale of multivariate characterization 132
7.4 Main steps of multivariate characterization 133
7.4.1 Compilation of chemical descriptors (Example: SOLVENT)....133
7.4.2 Derivation of principal properties (Example: SOLVENT) 133
7.4.3 Selection of a representative set of observations (Example: SOLVENT) 135
7.4.4 A short discussion 137
7.5 Main example: SURFACTANTS 137
7.5.1 Multivariate characterization of technical surfactant blends 138
7.5.2 Training set selection 139
7.5.3 Detergency measurements 139
7.5.4 Acute toxicity measurements 139
7.5.5 PLS regression between chemical descriptors (X) and detergency/toxicity (Y)....140
7.5.6 Predictions of response profiles (Y) for non-tested surfactants 143
7.5.7 Epilogue 143
7.6 Questions for Chapter 7 144
7.7 Summary and discussion 144

8 Multivariate calibration 147

8.1 Objective 147
8.2 Introduction....147
8.2.1 An etymological consideration 147
8.2.2 Calibration – What is it? 147
8.2.3 Limitations of traditional calibration 148
8.3 Multivariate calibration 149
8.3.1 Introduction to multivariate calibration 149
8.3.2 A geometric interpretation of multivariate calibration 150
8.3.3 Main steps of multivariate calibration 151
8.3.4 Some application areas....152
8.4 Example: SUGAR....153
8.4.1 Background 153
8.4.2 PCA of the spectral data (X) 153
8.4.3 PLS modeling of the entire dataset 155
8.4.4 PLS modeling of the reduced dataset 156
8.4.5 OPLS modeling of the reduced dataset 157
8.4.6 Summary of SUGAR application 157
8.5 A second example: SOIL_DNA....158
8.5.1 Background 158
8.5.2 PCA of the fingerprint data (X) 159
8.5.3 PCA of the chemical concentration data (Y) 160
8.5.4 PLS modeling 161
8.5.5 OPLS modeling 163
8.5.6 Summary of SOIL_DNA application 163
8.6 A third example: SAWDUST 163
8.6.1 Background 163
8.6.2 PCA of NIR data (X) 164
8.6.3 PLS modeling 165
8.6.4 OPLS modeling 165
8.6.5 Summary of SAWDUST application 166
8.7 Questions for Chapter 8 166
8.8 Summary and discussion 166
8.8.1 Two calibration situations 167

8.8.2 Reference samples with one or two constituents 167
8.8.3 Reference samples with multiple constituents 167
8.8.4 Multiple constituents varied via mixture design 167
8.8.5 Combined design in constituents and process conditions 168
8.8.6 Samples are collected from the system and analyzed by a reference method 168
8.8.7 Location-based sampling 168
8.8.8 Batch-based sampling 169

9 Multivariate process modeling 171

9.1 Objective 171
9.2 Introduction 171
9.3 Multivariate process modeling 172
9.3.1 Objectives 172
9.3.2 Process data 172
9.3.3 Applications 173
9.4 Example: CUPRUM 174
9.4.1 Quality control with a univariate index 174
9.4.2 A better approach - PCA 174
9.4.3 Zooming-in on highlighted samples 175
9.4.4 Contribution-based diagnostics 176
9.4.5 Plots of T^2 and DModX are more informative than TAI 177
9.4.6 Summary of CUPRUM application 178
9.5 A second example: PROC1A 178
9.5.1 Establishing an overview of the process data 178
9.5.2 Modeling the normal process behavior 179
9.5.3 Using the model to monitor the process 180
9.5.4 Identifying contribution from variables 180
9.5.5 Exploring time series properties of scores 181
9.5.6 Summary of PROC1A 182
9.6 A third example: SOVRING 182
9.6.1 An overview PCA model 182
9.6.2 A refined PCA model 183
9.6.3 PLS modeling of 12 factors, 6 expanded terms, and 6 responses 184
9.6.4 Evaluation of PLS model 184
9.6.5 Interpretation of PLS model 185
9.6.6 Response contour plotting 186
9.6.7 Using the full PLS model for process monitoring 187
9.6.8 Using the full PLS model for on-line predictions 188
9.6.9 Summary of SOVRING application 189
9.7 Questions for Chapter 9 189
9.8 Summary and discussion 189

10 Classification and discrimination 191

10.1 Objective 191
10.2 Introduction 191
10.2.1 Some typical classification situations 191
10.2.2 Approaches to class specification 192
10.3 Pattern recognition (PARC) 192
10.3.1 Level 1: Classification into one of a number of specified classes 193
10.3.2 Level 2: Level 1 plus the possibility of outliers 193
10.3.3 Level 3: Level 2 plus a response variable (y) 194
10.3.4 Level 4: Level 3, but with a matrix of responses (Y) 195
10.4 Example: IRIS 195
10.4.1 An overview PCA model 195
10.4.2 Fitting a new PCA model to the Versicolor and Virginica classes 196

10.4.3 Soft independent modeling of class analogy (SIMCA) 196
10.4.4 SIMCA of the IRIS dataset 197
10.4.5 The Coomans' plot: A powerful graphical display 199
10.4.6 Summary of the IRIS application 200
10.4.7 A recent extension to the IRIS dataset 200
10.4.8 Classification of the extended dataset into the overview PCA model 200
10.4.9 Classification of the Yellow Flags into the local Virginica model 201
10.4.10 Group contribution plot to compare two groups 202
10.4.11 Final summary of the IRIS application 202
10.5 Partial least squares discriminant analysis (PLS-DA) 203
10.6 Example: ARCHAE 204
10.6.1 An overview PCA model 204
10.6.2 SIMCA of archaeology data 205
10.6.3 PLS-DA of archaeology data 206
10.6.4 Summary of ARCHAE application 208
10.7 Questions for Chapter 10 208
10.8 Summary and discussion 208
10.8.1 How to formulate the classification problem 208
10.8.2 Main steps of SIMCA 209
10.8.3 When and when not to use PLS-DA 209
10.8.4 Regular and combined DModX´s 210
10.8.5 The relationship between the Coomans' plot and the Classification and Misclassification lists 211
10.8.6 Related methods 213

11 Identification of discriminating variables 215

11.1 Objective 215
11.2 Introduction 215
11.3 A tutorial example: Distilled Liquor 216
11.3.1 Introduction 216
11.3.2 Two-class discriminant modeling 216
11.4 Omics example: Poplar 218
11.4.1 Introduction 218
11.4.2 OPLS-DA models between control and treatment groups 219
11.4.3 The S-plots 220
11.4.4 Interpretation using the S-plots 221
11.4.5 Summary of the Poplar example 223
11.5 Main example: Wafers 224
11.5.1 Introduction 224
11.5.2 OPLS-DA model between the three chambers 224
11.5.3 Interpretation of the predictive components 225
11.5.4 Interpretation of the orthogonal components 227
11.5.5 Summary of the Wafer example 229
11.6 Questions for Chapter 11 230
11.7 Discussion 230

12 Transformation and expansion 233

12.1 Objective 233
12.2 Transformation 233
12.2.1 Introduction 233
12.2.2 Which transformations are useful for approaching normality? 234
12.2.3 Benefits of transformation 235
12.3 Example: NEURO 236
12.4 Expansion of the X-matrix 239
12.4.1 Introduction 239

12.4.2 Implicit non-linear latent variable regression, INLR ... 240
12.5 Example: SOVRING ... 240
12.6 Questions for Chapter 12 ... 242
12.7 Summary ... 242

13 Centering and Scaling 243

13.1 Objective ... 243
13.2 Introduction ... 243
13.2.1 Scaling to unit variance (UV-scaling) ... 243
13.2.2 No scaling and Pareto scaling ... 244
13.2.3 Set-point centering and control/action-limit scaling ... 244
13.2.4 Inclusion/exclusion of variables is a way of scaling data ... 244
13.3 Example: LOWARP ... 244
13.4 Example: SUGAR ... 245
13.5 Related scaling procedures ... 247
13.5.1 Block-scaling ... 247
13.5.2 Scaling, but no mean-centering ... 249
13.5.3 Double-centering ... 251
13.6 Trimming and Winsorizing ... 252
13.7 Questions for Chapter 13 ... 253
13.8 Summary ... 254

14 Signal correction and compression 255

14.1 Objective ... 255
14.2 Introduction ... 255
14.3 Methods for signal correction and compression ... 256
14.3.1 Orthogonal signal correction, OSC ... 257
14.3.2 Multiplicative signal correction, MSC ... 258
14.3.3 Standard normal variate (SNV) correction ... 258
14.3.4 First, second and third order derivation ... 258
14.3.5 Chained filters ... 259
14.3.6 Wavelet analysis ... 259
14.3.7 Main steps in data compression ... 260
14.4 An example on chained filters ... 261
14.5 Questions for Chapter 14 ... 262
14.6 Summary and discussion ... 262

15 MSPC 265

15.1 Objective ... 265
15.2 Introduction ... 265
15.3 Statistical process control (SPC) ... 267
15.3.1 Basic assumption of SPC ... 267
15.3.2 Approach to detect the occurrence of a special event ... 267
15.3.3 Scope of SPC ... 268
15.3.4 Transition from SPC to MSPC ... 269
15.4 The Shewhart chart ... 270
15.4.1 Construction in univariate application ... 270
15.4.2 Multivariate Shewhart charts ... 270
15.4.3 Subgrouping ... 272
15.5 The CuSum chart ... 273
15.5.1 A cumulative sum of differences between target and measured values ... 273
15.5.2 Ways of implementing CuSum ... 273
15.5.3 Multivariate CuSum charts ... 274
15.6 The EWMA chart ... 275

15.6.1 An approach towards exploring process memory 275
15.6.2 Ways of implementing EWMA 275
15.6.3 Multivariate EWMA charts 275
15.6.4 EWMA and its relation to Shewhart and CuSum charts 276
15.7 Example: PROC1A 277
15.7.1 A score scatter plot shows direction of change 277
15.7.2 A multivariate T^2 control chart shows magnitude of deviation (but not direction) 277
15.7.3 Important scores can be charted alone 278
15.7.4 Real-time process monitoring using SIMCA-online 279
15.7.5 Summary of MSPC of PROC1A dataset 282
15.8 Questions for Chapter 15 282
15.9 Summary and discussion 282
15.9.1 The basic control charts 282
15.9.2 Control charts limits 283
15.9.4 SPC/MSPC and relation to feedback process control 284
15.9.5 Alternative approaches to process monitoring and control 284

16 BSPC 287

16.1 Objective 287
16.2 Introduction 287
16.2.1 A finite duration process 287
16.2.2 Batch processes form three-way data tables 288
16.2.3 Multivariate analysis is crucial in batch monitoring 289
16.2.4 Objectives of multivariate batch modeling 289
16.3 A multivariate approach to BSPC 290
16.3.1 Example dataset: BAKER'S YEAST 290
16.3.2 Two types of batch models 291
16.4 The batch evolution model (EXAMPLE: BAKER'S YEAST) 291
16.4.1 Unfolding of data preserving the variable direction 291
16.4.2 Creating a BEM of the 16 accepted batches 292
16.4.3 Rearrangement of scores to give batch control charts 292
16.4.4 Batch control charts of the BEM 293
16.4.5 Monitoring the evolution of new batches 294
16.4.6 Finding variables contributing to deviation from batch normality 295
16.4.7 Summary of batch evolution modeling 296
16.5 The batch level model (EXAMPLE: BAKER'S YEAST) 296
16.5.1 Unfolding of data preserving batch direction 296
16.5.2 Computing a BLM on the 16 reference batches 297
16.5.3 Model interpretation 297
16.5.4 Classification of new batches 298
16.5.5 Summary of batch level modeling 299
16.6 A second batch example: Wafers 299
16.6.1 Background and arrangement of data 299
16.6.2 Objectives of the BSPC study 299
16.6.3 Batch evolution model 299
16.6.4 Use of BEM: Classification of test set batches 300
16.6.5 Batch level model 301
16.6.6 Use of BLM: Validation using the test set 302
16.6.7 Summary of Wafers application 303
16.7 Questions for Chapter 16 303
16.8 Summary and discussion 304
16.8.1 BSPC – How we do it 304
16.8.2 Some advantages of the BSPC strategy 304
16.8.3 Going online 304
16.8.4 Related approaches 305
16.8.5 Batch duration and synchronization 305

17 Multivariate time series analysis 307

17.1 Objective ... 307
17.2 Outline of Chapter 17 ... 307
17.3 Introduction to time series data ... 307
17.3.1 Definition of time series data ... 307
17.3.2 Multivariate time series data ... 308
17.3.3 Objectives of multivariate time series analysis (MTSA) ... 308
17.3.4 Transfer functions and PLS time series analysis ... 308
17.4 Review of time series and their properties ... 308
17.4.1 Stationary stochastic models ... 308
17.4.2 Autocorrelation and partial autocorrelation of the AR, MA, and ARMA models ... 309
17.4.3 Non-stationary stochastic models ... 309
17.4.4 Identifying the time series model ... 310
17.4.5 Modeling multivariate time series data ... 310
17.5 The transfer function and PLS time series analysis ... 310
17.5.1 Introduction ... 310
17.5.2 The FIR Model ... 311
17.5.3 Transfer function models ... 311
17.5.4 Identification of transfer function model ... 312
17.6 MTSA and lagging of latent variables (Example: CUPRUM) ... 312
17.6.1 PCA of original CUPRUM data ... 312
17.6.2 MTSA of the t_1-score ... 312
17.6.3 MTSA of the t_2-score ... 314
17.6.4 Summary of MTSA of CUPRUM data ... 315
17.7 MTSA using PLS (Example: KAMYR) ... 316
17.7.1 The KAMYR digester process ... 316
17.7.2 The autocorrelation function ... 317
17.7.3 The cross-correlation function ... 318
17.7.4 PLS modeling of 21 X-variables and the Kappa number ... 318
17.7.5 PLS time series analysis: Lagging of variables ... 319
17.7.6 PLS finite impulse response (FIR) model of Kappa number ... 320
17.7.7 PLS transfer function model of Kappa number ... 320
17.7.8 Summary of MTSA of the KAMYR data ... 321
17.8 Questions for Chapter 17 ... 322
17.9 Summary and discussion ... 322

18 Process Analytical Technology (PAT) and Quality by Design (QBD) 323

18.1 Objective ... 323
18.2 Introduction ... 323
18.2.1 The PAT initiative ... 323
18.2.2 What are the benefits of using DOE? ... 324
18.2.3 QBD and Design Space ... 325
18.2.4 MVDA/DOE is needed to accomplish PAT/QBD in Pharma ... 325
18.2.5 Organization of Chapter 18 ... 326
18.3 A process optimization case study: SOVRING ... 326
18.3.1 The SOVRING dataset ... 327
18.3.2 PLS modeling of 85-samples SOVRING subset ... 327
18.3.3 Summary of application ... 330
18.4 PAT: Goals, components, pre-requisites and levels ... 330
18.4.1 Goals of PAT according to FDA ... 330
18.4.2 Results of PAT ... 331
18.4.3 Pre-requisites of PAT ... 332
18.4.4 Four levels of PAT ... 333
18.5 A PAT example: Binary Powder II ... 335

18.5.1 Background to example and objectives 335
18.5.2 Primary dataset 335
18.5.3 Monitoring dataset....336
18.5.4 Evaluation of raw data....336
18.5.5 PCA for data overview 337
18.5.6 PLS model based on the training set 339
18.5.7 Application to the monitoring dataset....341
18.5.8 SNV pre-processing for improved predictive power 343
18.5.9 OPLS for enhanced interpretability 344
18.5.10 Discussion of example 346
18.6 Combining the concepts: The Novartis PAT example 348
18.6.1 Background 348
18.6.2 Workflow to synchronize data 349
18.6.3 Nine base level batch models 350
18.6.4 Top level batch model 350
18.6.5 Discussion and Epilogue....351
18.7 Questions for Chapter 18 351
18.8 Summary and discussion 351

19 Hierarchical modeling 355

19.1 Objective 355
19.2 Introduction....355
19.3 Process example: HI-PROC 356
19.3.1 Base level block formation 356
19.3.2 Base level model structure....357
19.3.3 Summarizing input and feed (1st base level model)....357
19.3.4 Summarizing reaction conditions (2nd base level model)....357
19.3.5 Summarizing the purification step (3rd base level model) 358
19.3.6 Summarizing the less important Y-variables (4th base level model) 358
19.3.7 Top level model: PSV of the entire process....359
19.3.8 Monitoring of the prediction set data....361
19.3.9 Discussion of HI-PROC example....362
19.4 Spectroscopy example: Carrageenan....362
19.4.1 Background and objective....362
19.4.2 Mixture design to get representative Y-data in the training and test sets....363
19.4.3 The three blocks of spectral data....363
19.4.4 Separating Y-predictive and Y-orthogonal spectral variation....363
19.4.5 Discussion of example 367
19.5 Questions for Chapter 19 367
19.6 Summary and discussion 368
19.6.1 Zoom-in/zoom-out capability....368
19.6.2 Focus on the “right” variables at each level 368
19.6.3 An alternative to variable selection 368
19.6.4 An alternative to block-scaling 369
19.6.5 Filtering of data 369
19.6.6 Missing data and outlier handling 370
19.6.7 Principal component regression, PCR 371
19.6.8 Related multi-block techniques 371

20 Non-linear PLS modeling 373

20.1 Objective ... 373
20.2 Introduction ... 373
20.2.1 General considerations ... 373
20.2.2 Degrees of non-linearity ... 374
20.2.3 Approaches to diagnosing and handling non-linearity ... 375
20.2.4 Non-linear PLS and other non-linear modeling tools ... 375
20.3 Binning and expansion of X-variables (GIFI-PLS) ... 377
20.3.1 Introduction to GIFI-PLS ... 377
20.3.2 Main steps of GIFI-PLS ... 377
20.3.3 Properties of GIFI-PLS ... 378
20.4 QSAR Example: ELASTASE (Peptide QSAR) ... 378
20.4.1 The data ... 378
20.4.2 Reference PLS model ... 379
20.4.3 GIFI-PLS model ... 379
20.4.4 Full PLS model ... 380
20.4.5 External predictivity assessment ... 380
20.4.6 Summary ... 380
20.5 Process Example: SimCODM ... 380
20.5.1 Introduction to SimCODM ... 380
20.5.2 Reference PLS model ... 381
20.5.3 PLS modeling using original and binned process variables ... 382
20.5.4 Where lies the non-linearity? ... 382
20.5.5 Summary ... 384
20.6 Questions for Chapter 20 ... 384
20.7 Summary and discussion ... 385

21 Hierarchical cluster analysis, HCA 387

21.1 Objective ... 387
21.2 Introduction ... 387
21.2.1 What is a dendrogram? ... 388
21.2.2 Bottom-up and top-down approaches ... 389
21.3 HCA in the SIMCA software ... 389
21.3.1 Recommended procedure involves scores or loadings ... 389
21.3.2 Choice of linkage criterion ... 389
21.3.3 Facilitated interpretation by connecting the dendrogram plot to other plots ... 390
21.4: An example: A clustered process ... 390
21.4.1 Introduction to the dataset ... 390
21.4.2 Outlier detection and removal ... 391
21.4.3 Updated PCA overview model ... 391
21.4.4 HCA in the six principal component score vectors ... 392
21.4.5 Evaluation of the resulting cluster pattern ... 392
21.4.6 Summary of the example ... 393
21.5 Questions for Chapter 21 ... 394
21.6 Summary and discussion ... 394

22 PLS-Trees 397

22.1 Objective ... 397
22.2 Introduction ... 397
22.2.1 General considerations ... 397
22.2.2 Latent variable-based approaches to clustering ... 398
22.2.3 A fast, robust and flexible approach ... 398
22.3 Theory of the PLS-Tree method ... 399
22.4 An example: DataMining-QSAR ... 400

22.4.1 Background to the dataset 400
22.4.2 Data-mining via top-down clustering 401
22.4.3 Zooming-in on stable structure-activity relationships 401
22.4.4 Summary of the example 403
22.5 Questions for Chapter 22 403
22.6 Summary and discussion 403

Appendix I: Model derivation, interpretation, and validation 407

AI.1 Objective 407
AI.2 Introduction 407
AI.2.1 Main steps of MVDA 407
AI.2.2 The iterative nature of MVDA 408
AI.3 Evaluation and pre-processing of raw data 409
AI.3.1 Main steps 409
AI.3.2 What should we look for in univariate statistics? 410
AI.3.3 Efficacious plots 410
AI.4 Model derivation and interpretation 411
AI.4.1 Main steps 411
AI.4.2 PCA for data overview 411
AI.4.3 PLS for quantitative modeling and predictions 414
AI.4.4 OPLS for enhanced interpretation 417
AI.4.5 O2PLS for data integration 418
AI.5 Model validation and use of model 419
AI.5.1 Main steps 419
AI.5.2 Why is predictive validation necessary at all? 419
AI.5.3 Internal validation (cross-validation) 420
AI.5.4 Response permutation and cross-validation 421
AI.5.5 External validation 423
AI.5.6 Predictions of new observations (Use of model) 424
AI.6 Questions for Appendix I 424
AI.7 Summary and discussion 424
AI.7.1 MVDA is an iterative procedure 424

Appendix II: Statistics 425

Introduction 425
Fit methods background 425
PCA - Principal Components modeling 425
Number of model dimensions (A) 426
Reference PCA 426
PLS - Partial Least Squares Projection to Latent Structures modeling 426
Number of model dimensions (A) 427
Reference PLS 427
OPLS/O2PLS - Orthogonal PLS modeling 427
OPLS - Orthogonal PLS modeling 427
O2PLS - Orthogonal PLS modeling 428
Conventional PLS compared with OPLS and O2PLS 429
PLS 429
OPLS 429
O2PLS 429
Combination with the hierarchical approach 429
Y-Related Profiles 430
References OPLS and O2PLS 430
Cluster Analysis (CA), dendrograms, Hierarchical CA (HCA), PLS-Tree 430
Hierarchical Cluster Analysis - HCA 431
Ward clustering 431
Single linkage 431

PLS-Trees ... 431
Calculation of PLS-Trees ... 432
Vectors available in SIMCA ... 433
Variables and Scores vectors ... 433
Observations and Loadings vectors ... 438
Function of component vectors ... 443
Function of lags vectors ... 445
Aligned Vectors ... 445
Batch Vectors ... 446
Formulas and descriptions ... 448
Prediction vectors ... 448
R2Y, R2X, R2Yadj, R2Xadj ... 448
R2V and R2Vadj ... 448
Q2 and Q2V ... 448
Q2(cum) and Q2V(cum) ... 449
Leverages ... 449
RSD of observations and variables ... 449
Relevance of variables ... 449
Modeling power - MPowX ... 450
Hotelling's T2 ... 450
Ellipse in score plot ... 450
Missing values correction factor ... 450
Missing values and PCA ... 450
Score and loading vectors ... 451
T scores (X): t1 vs. t2, ... 451
U scores (Y): u1 vs. u2, ... 451
T vs. U scores (X&Y): t1 vs. u1, ... 451
P loadings (X): p1 vs. p2,... ... 451
W loadings (X): w1 vs. w2, ... 451
W* loadings (X): w*1 vs. w*2, ... 451
Relation of W* to the PLS regression coefficients ... 452
C loadings (Y): c1 vs. c2, ... 452
WC or W*C loadings (X&Y): wc1 vs. wc2, ... 452
TPS scores (X): tPS1 vs. tPS2, ... 452
Distance to the model ... 452
Absolute distance to the model of an observation in the workset - DModX, DModY(Absolute) ... 452
DModX absolute ... 452
DModY absolute ... 453
DModY absolute for OPLS/O2PLS ... 453
Normalized distance to the model of an observation in the workset - DModX, DModY (Normalized) ... 453
DModX normalized ... 453
DModY normalized ... 453
Absolute distance to the model of new observations in the predictionset - DModXPS, DModYPS (Absolute) ... 453
Normalized distance of a new observation to the model - DModXPS, DModYPS (Normalized) ... 454
Distance to the model augmented - DModXPS+ (Absolute) or (Normalized) ... 454
MPow weighted distance to the model - DModX, DModXPS, DModXPS+ (Abs/Norm and weighted) ... 454
Critical distance to the model ... 454
Membership significance level ... 455
Variable importance, VIP ... 455
Standard Error and Jack-knifing uncertainties ... 456
Standard error according to Höskuldsson - Serr ... 456

Jack-knifing (JK) uncertainties 456
Reference Jack-knifing 456
Coefficients 456
Scaled and centered coefficients - CoeffCS 457
Coefficients MLR - CoeffMLR 457
Coefficients unscaled - Coeff 457
Rotated coefficients 458
Reference rotated coefficients 458
Coefficients with Y unscaled uncentered, X unscaled centered - CoeffC 458
RMSEE, RMSEP and RMSECV 458
Transform page criteria 459
Automatic transformation criteria 459
Scaling 459
Autoscaling 459
Block scaling 459
Scaling weight calculation 460
Scaling of expanded terms 460
Computation 460
Scaling and centering of the expanded term after computation 460
Scaling weights for transformed variables 460
Scaling of variables with classes 460
Scaling of lagged variables 461
Scaling after changing the observation selection 461
Cross validation 461
Cross validation for PCA 461
PRediction Error Sum of Squares - PRESS 461
Cross validation for PLS, OPLS and O2PLS 461
PRediction Error Sum of Squares - PRESS 461
PLS Specific 462
OPLS and O2PLS Specific 462
Reference cross validation 462
Cross validation rules - Significant component 462
Rule 1: R1 462
Rule 2: R2 for PCA 462
Rule 2: R2 for PLS, OPLS and O2PLS 463
Rule 3: U and R5 463
Cross validation rules - Non Significant component 463
N3 463
N4 463
NS 463
Cross validation rules for batches 463
Rule 1: RB1 463
Rule 2: RB2 464
NB1 - stopping rule 464
NS - Not Significant 464
PLS time series analysis 464
References PLS time series 464
CV-ANOVA 464
The ANOVA table 465
References CV-ANOVA 466
Fisher's Exact test 466
References Fisher's Exact test 466
Control Chart statistics 466
Nomenclature and notation in Control Charts 466
Target and Standard deviation 467
S-plot 468
Reference S-plot 468

Observation Risk 468

References **469**

References for MVDA book 469

Index **491**

0 Preface

0.1 Why MultiVariate Data Analysis (MVDA) ?

To understand the world around us, as well as ourselves, we need to measure many things, many variables, many properties of the systems and processes we investigate. Hence, data collected in science, technology, and almost everywhere else are multivariate, with multiple variables measured on multiple samples or at multiple time points.

Multivariate data, accurately measured on intelligently selected observations and variables, contain much more information than univariate data, and hence an adequate multivariate characterization is a necessary first step in their investigation. This is so for both intellectual and practical purposes, i.e., in both basic research and applied technology.

However, in order to use a multivariate dataset to obtain insight into the system studied, it is not enough to just look at the data table. Rather, the data must be analyzed so that the information in the data is expressed in a comprehensible way, for instance as a graph, or as a few information-rich parameters. After this multivariate analysis, the results are interpreted, i.e., related to the objectives of the investigation and the scientific-technical context.

0.2 Projection Methods (Principal Component Analysis, PCA, and Projections to Latent Structures, PLS)

This book describes a remarkably simple approach to multivariate analysis based on so called projection methods. This approach represents the "cases" as a swarm of points in a K-dimensional space (K = number of variables), and then projects the point swarm down onto a lower-dimensional plane or hyper-plane. The coordinates of the points on this hyper-plane provide a compressed representation of the observations, and the direction vectors of the hyper-plane provide a corresponding representation of the variables.

The projection approach can be adapted to a variety of data-analytical objectives, i.e., (i) summarizing and visualizing a dataset, (ii) multivariate classification and discriminant analysis, and (iii) finding quantitative relationships among the variables. This applies to any shape of multivariate data, with many or few variables, many or few observations, and complete or incomplete data tables (missing data). In particular, projections handle data matrices with more variables than observations very well, and the data can be noisy and highly collinear.

With small modifications, projection methods can be made robust to outliers, deal with non-linear relationships, and adapt to drift in multivariate process data.

0.3 Recognition of differences between multi- and megavariate analysis

Traditionally, multivariate analysis has meant multiple linear regression (MLR), linear discriminant analysis (LDA), canonical correlation (CC), factor analysis (FA) and principal component analysis (PCA) applied to independent variables, i.e., cases where the variance-covariance matrix, **X'X,** has full rank and is fairly well conditioned.

In chemical and engineering practice, however, we often assume that our systems are driven by inherent, latent, variables, which are few compared with the number of observed

variables, *K*. Hence, the data and variance-covariance matrices, **X** and **X'X**, do not have full rank, but rather have the practical rank *A*. Methods used in this situation are PCA for projecting **X** down onto a few latent variables, SIMCA and PLS-DA for classification, and PLS and PCR for latent variable regression. The latent variable models are philosophically different in objectives and formulation from the traditional multivariate models with independent variables.

To distinguish between these two types of situation (with related data, models, and data analytical methods), we shall often refer to the latter as *megavariate* (from Greek "large", e.g., megalopolis). Megavariate analysis models data in terms of multiple latent variables, to give results that are multivariate. Since this is a new nomenclature, however, it will not be used consistently all through the book, but rather in situations where we feel a need to distinguish between the situation where **X** is full rank and the more common megavariate situation where **X** has a much lower rank than both the number of variables (*K*) and the number of observations, *N*. Megavariate data analytical problems are increasingly encountered in e.g. systems biology and life sciences [Grainger, 2003].

0.4 Problems areas where MVDA is useful

The flexibility of projection methods have made them useful also for the analysis and modeling of messy and complicated data; these methods are increasingly used in a wide range of industrial applications in PARD (processes, administration, research, and development). These include process monitoring and early fault detection and classification, process analytical technology (PAT), data mining and integration, quality control, and relationships between chemical composition/structure and chemical/biological properties. Multivariate calibration, which models the relationship between a multitude of signals measured on various samples and the concentrations of constituents of these samples, is another show-case of MVDA and projections.

Application areas of MVDA:

- Process monitoring, and early fault detection and classification
- Process analytical technology (PAT)
- Quality control
- Data mining and integration
- Composition – property relationships
- Structure – activity (and property) relationships
- Multivariate calibration
- Multivariate characterization
- Image analysis

Industrial sectors using MVDA of data from PARD:

- Basic chemicals
- Petrochemicals
- Polymers
- Plastics
- Fibers
- Resins, Paints, Adhesives
- Automobiles
- Pharmaceuticals
- Biotech
- Paper and Pulp

- Food and Beverages
- Mining and Prospecting
- Metals and Materials
- Semiconductors
- Telecommunications
- Marketing data analysis
- Consumer science

0.5 Objectives of the book

The purpose of this book is to provide a "minimal" and practical approach to MVDA that is useful with almost all multivariate data, and to show how this approach can be used in science and technology for a wide variety of applications. In particular, we want to show how the extensive information contained in multivariate data can be expressed in terms of plots and lists of parameters resulting from a multivariate analysis. In turn, these results help us improve processes and products, as well as improve the efficiency of research and development. In industry this leads to great savings as well as increased income from higher yields and improved quality. In both industry and academia this also leads to an improved understanding of the systems under investigation. An illustration of the industrial potential of MVDA is given in Figure 0.1.

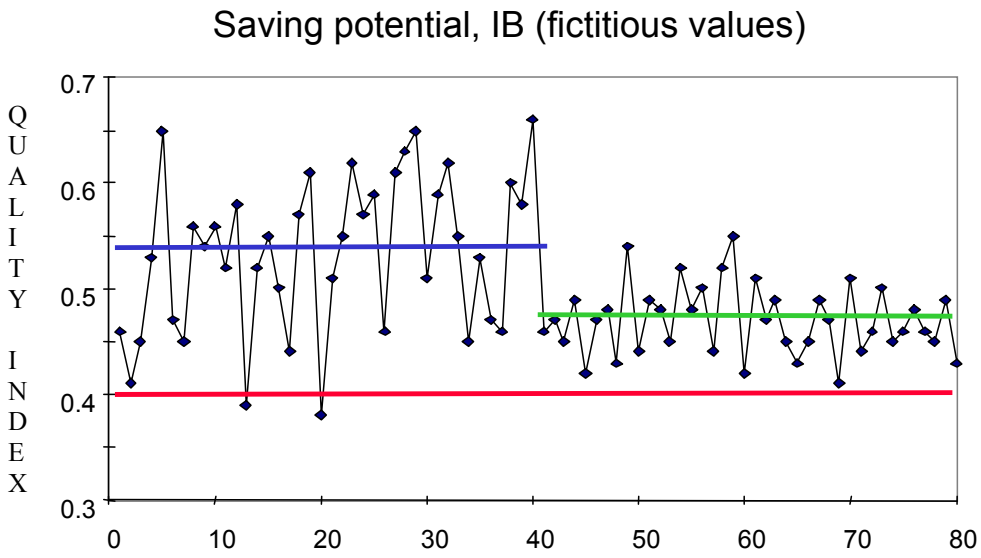


Figure 0.1: An illustration of the potential of MVDA to achieve savings and increase income. The example relates to a quality parameter. By using MVDA for on-line process modeling and monitoring it was possible to reduce product variability and enhance product quality. This also resulted in early fault detection and fewer process shut-downs, i.e., in increased productivity.

0.6 Organization of the book

The MVDA book comprises 22 chapters, two appendices and a references list. Appendix I is intended to serve as a flow-chart for MVDA, which can be consulted for guidance and overview by a user perhaps employing MVDA on a less regular basis. Appendix II is the statistical appendix.

Below, we give a short description of each segment of the book.

0.6.1 Section I, Chapters 1 – 6

The first section (Chapters 1 – 6) describes basic concepts and principles for multivariate data analysis by means of projection methods. Separate chapters are devoted to the description of principal component analysis (PCA), projections to latent structures (PLS), OPLS and O2PLS. Section I is intended for readers who only want a general overview of the field of MVDA by means of projection methods.

0.6.2 Section II, Chapters 7 – 11

The second section (Chapters 7 – 11) provides a thorough introduction to five important applications of MVDA. These five application areas are (i) multivariate characterization, (ii) multivariate calibration, (iii) multivariate process modeling and monitoring, (iv) multivariate classification and discriminant analysis, and (v) identification of discriminating variables. Section II is intended for new users of MVDA who want a broader perspective of fundamental applications of PCA, PLS, OPLS and O2PLS.

0.6.3 Section III, Chapters 12 – 14

In MVDA, data preparation and standardization is an important area. Section III of the book (Chapters 12 – 14) overviews basic pre-processing methods for (i) transformation, (ii) scaling, and (iii) correction and compression of multivariate data. The overall goal of data pre-processing is to shape the data to facilitate the subsequent data analysis. The third section of the book is intended for users of MVDA with some experience.

0.6.4 Section IV, Chapters 15 – 18

In Section IV (Chapters 15 – 18), we move into a more specific and dedicated treatment of process modeling and monitoring. The objectives of Section IV are to highlight methods for (i) multivariate statistical process control, MSPC, (ii) batch statistical process control, BSPC, (iii) multivariate time series analysis, MTSA, and (iv) process analytical technology (PAT). Section IV addresses users with some experience of MVDA from the process industrial sector.

0.6.5 Section V, Chapters 19 – 22

Section V (Chapters 19 – 22) provides material on method extensions, namely (i) hierarchical modeling, (ii) non-linear regression modeling, (iii) hierarchical cluster analysis and (iv) PLS-Trees. Section V is useful for experienced users that may want to use more advanced data analytical tools.

1 Introduction

1.1 Objective

This chapter gives a general introduction to multivariate data analysis and three basic problem types, (i) overview of data, (ii) classification/discrimination, and (iii) regression modeling. Two examples of each problem category will be given.

1.2 Introduction

In this book, we shall study methods for extracting information from large or small tables of data (see Figure 1.1). This is called multivariate data analysis, or MVDA for short. Tables of data are abundant in all parts of research and development, both in academia and industrial R&D, as well as in industrial manufacturing. This is due to the rapid instrumentalization of science and technology. Masses of data are produced by measuring many variables on ensembles of chemical samples, or by frequently recording many signals from an industrial process in order to track its behavior. The result is summarized in Figure 1.2.

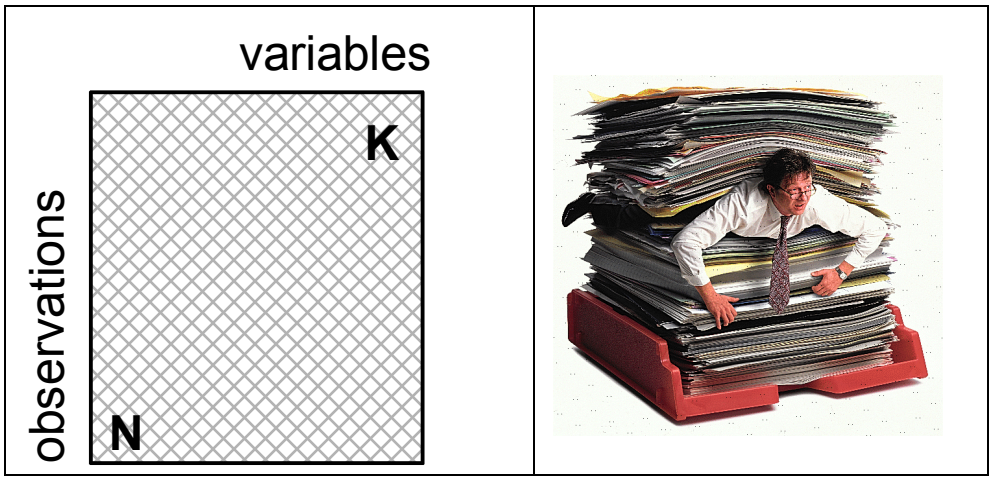


Figure 1.1: (left) A data table with N observations and K variables. The observations are sometimes also called objects, samples, cases, or items. The variables are the measurements that are made in order to capture the properties of the observations. Figure 1.2: (right) Masses of data require efficient data analytical tools to handle the information in the data.

The data explosion necessitates the use of appropriate analytical tools for extracting meaningful information from the large amount of raw data. It is no longer efficient to analyze data by simply looking at them (Figure 1.2) or by plotting them in simple graphs. More sophisticated, computer-based methods are needed if the data analysis is to be accomplished within a reasonable time.

The question is *how do we analyze these heaps of data*? This book will describe two multivariate projection methods which are useful: principal component analysis (PCA) [Jackson 1991] and partial least squares projections to latent structures (PLS) [Wold, *et al.*, 1984]. These methods are very easy to understand and to use.

1.3 Pattern recognition

In chemistry and chemical engineering, the term pattern recognition (PARC) is often used as a synonym for MVDA to indicate how multivariate data analysis finds the typical "data pattern" for one or several classes of observation [Wold, *et al.*, 1984]. The "pattern" of one class provides information about the relations between the observations within the class: which are close and which are far away from each other and which are "non-typical" outliers. Information is also obtained about similarities and dissimilarities among the variables.

If the "patterns" differ between classes, they can then be used to assign new observations to the classes, to classify new observations on the basis of the similarity between their data and the "class patterns". PARC can be generalized to the problem of finding patterns that express relations between blocks of variables measured on the same set of observations. In the simplest case there are two blocks, X and Y. This is also a generalization of regression and correlation. With combinations of classification and block-relations, most questions put to a data table acquired in research, development, and production, can be answered. Thus PARC is an empirical but general approach to the analysis of multivariate data, empirical in the sense that few fundamental assumptions or models are needed to perform the analysis.

1.4 Basic types of data analytical questions

MVDA provides a toolbox of flexible and versatile data analytical tools. There are three basic problem types to which these multivariate tools can be applied (Figure 1.3): (i) overview of a data table, (ii) classification and/or discrimination among groups of observations, and (iii) regression modeling between two blocks of data (X and Y). These three basic types of question also reflect the major stages of multivariate analysis. We often start with no specific question and desire only a brief summary. Then we may proceed to classification and end with quantitative modeling.

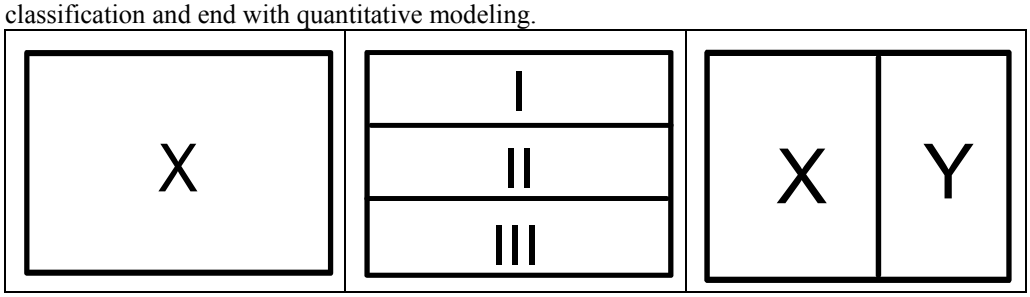


Figure 1.3: MVDA provides tools for three basic types of data analytical problem, namely overview (left), classification & discrimination (middle), and regression modeling between two blocks of data (right). In the overview stage, PCA is used and no distinction is made either between groups of observations or between categories of variables. PCA is also used at a later stage to make separate models of defined classes, with the aim of predicting class membership of additional observations, not previously considered in the data analysis. In order to link two blocks of variables, often referred to as X and Y, to each other, PLS is an appropriate method. PLS makes it possible to connect many X´s and many Y´s at the same time, thus enabling response profiles of the observations to be modeled and predicted.

1.4.1 Data overview

In the early stages of a project, when little is known about a problem, one often requires a simple overview of the information in a data table. Such an overview can be obtained with principal components analysis (PCA) [Jackson, 1991; Wold, *et al.*, 1984]. PCA produces a summary showing how the observations are related and if there are any deviating observations or groups of observations in the data. Of particular interest in process data analysis is the ability of PCA to uncover both smooth time trends and sudden shifts in the data. In addition, with PCA we also gain an understanding of the relationships among the variables: which variables contribute similar information to the PCA model, and which provide unique information about the observations. PCA describes the correlation structure in X.

1.4.2 Classification and discrimination

Often an initial PCA of a dataset reveals groupings among the observations. Two or three major groups of observations is not uncommon. This may indicate the need for further PCA modeling of each such group ("class") in order to fine-tune the analysis and to understand the characteristic features of the different groups. Subsequently, it is possible to use the separate class models for classification. By "classification" we mean the use of the established class models for classifying new observations. This results in a predicted class membership for these new observations based on their fit to the respective class model. Sometimes the class membership of an unknown observation is clear-cut, but sometimes the assignment is more ambiguous, or even impossible. The latter situation arises when a new sample does not fit any of the established class models. Often this kind of deviating observation, which has properties clearly different from the hitherto modeled ones, might be the interesting case for further investigation.

1.4.3 Regression modeling

The last stage of data analysis is regression modeling between two blocks of data, usually denoted X and Y, with the aim of predicting Y from X for new observations. This kind of modeling is accomplished with the PLS method [Höskuldsson, 1996; Wold, *et al.*, 1984; Wold, *et al.*, 2000; Wold, *et al.*, 2001], which may be thought of as the regression extension of PCA. In this book we shall call the X-variables *factors* or *predictors* and the Y-variables *responses*. In process modeling and monitoring the factors are the signals which are sampled frequently, at regular time intervals, to monitor the status of the process, whereas the responses mirror properties like yield or quality of a product, and which are measured less frequently. The responses are also often laborious, time-consuming, and expensive to measure compared with the factors.

In PLS modeling of complex process data, a major aim is to accomplish fast, accurate, and quantitative predictions of complex responses (e.g. quality of product, taste of wine, impurity of copper.....) based the collected body of X-data. As a consequence, this stage could equally well be termed *quantification and prediction*. Furthermore, with appropriate data and a workable PLS model (there are tools to test this), it is possible to find out how the factors influence the responses, how the responses correlate with each other, and how to adjust the factors to get a desired profile of the responses.

In the next three sections we will present some examples pertaining to each of the three major data analytical problems. These examples will be used as illustrations in later chapters.

1.5 Illustrations of problem types – Overview of a dataset

1.5.1 A getting started dataset (FOODS)

One dataset which we will study at the beginning of the book is the FOODS dataset. The aim of this investigation was to see how food consumption patterns varied between different European countries. The relative consumption of 20 food items was compiled for 16 countries. The dataset is listed in Table 1.1. The values range between 0 and 100% and a high value corresponds to a high consumption. Note that there are three missing data elements (observations 11, 14 and 15 in columns 4, 5 and 19 of Table 1.1), but this is not a problem since multivariate methods can handle missing data.

Table 1.1: The relative consumption of 20 food products across 16 European countries. Each entry shows the percentage of households that normally use each food item.

Primary ID	ONAM	Gr_Coffe	Inst_Coffe	Tea	Sweetner	Biscuits	Pa_Soup	Ti_Soup	In_Potat	Fro_Fish	Fro_Veg	Apples	Oranges	Ti_Fruit	Jam	Garlic	Butter	Margarine	Olive_Oil	Yoghurt	Crisp_Bread
1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
2	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
3	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
4	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
5	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
6	Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
7	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
8	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
9	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
10	Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
11	Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
12	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
13	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
14	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
15	Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
16	Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

The aim of this study was to examine the similarities and differences between the 16 countries with respect to their usage of the different food products. PCA creates a condensed summary of the table, which can be analyzed graphically by means of two plots, the *score* plot (Figure 1.4) and the *loading* plot (Figure 1.5). The score plot is a summary of the relationships among the observations (European countries) and the loading plot a similar summary of the variables (food products). The loading plot is a means to interpret the patterns seen in the score plot. The two plots are complementary and superimposable, and a direction in one plot corresponds to the same direction in the other plot.

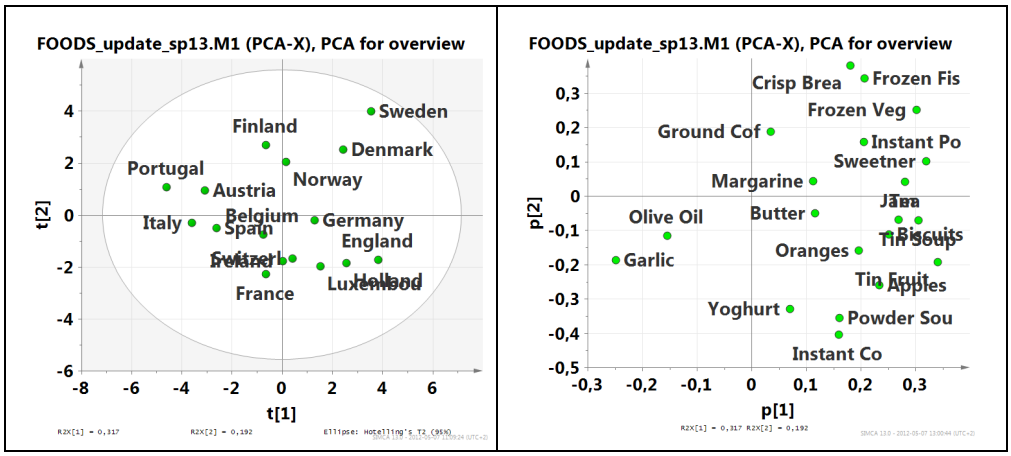


Figure 1.4: (left) PCA score plot of the FOODS dataset. Each plot mark corresponds to an observation (country).
Figure 1.5: (right) PCA loading plot of the FOODS dataset. Each plot mark represents a variable (food item).

In this case the score plot is a map of the relationships among the 16 countries. We recognize that the four Nordic countries are situated close together (see upper right-hand quadrant). This indicates that they share many food consumption habits. To understand why this is the case, we may look at the loading plot. One reason for the Nordic grouping has to do with high consumption of products like crisp bread, frozen fish, and frozen vegetables. This is inferred from the fact that these three variables are found far from the plot origin, and in the same part of the plot (upper right-hand quadrant) as the countries themselves are located in the score plot. Furthermore, we can see that three of the southern European countries, Portugal, Italy, and Spain, form a group together with Austria. According to the loading plot, in these four nations the inhabitants indulge in garlic and olive oil, and eat small amounts of tinned fruit, tinned soup, and sweetener.

Additionally, PCA enables an inspection of the model residuals in terms of the DModX parameter. DModX is an acronym of *d*istance to the *mod*el in the X-data, and indicates how well an observation fits the PCA model. The DModX plot of the FOODS dataset is given in Figure 1.6. We can see that all countries fit the model well, i.e., there are no outliers.

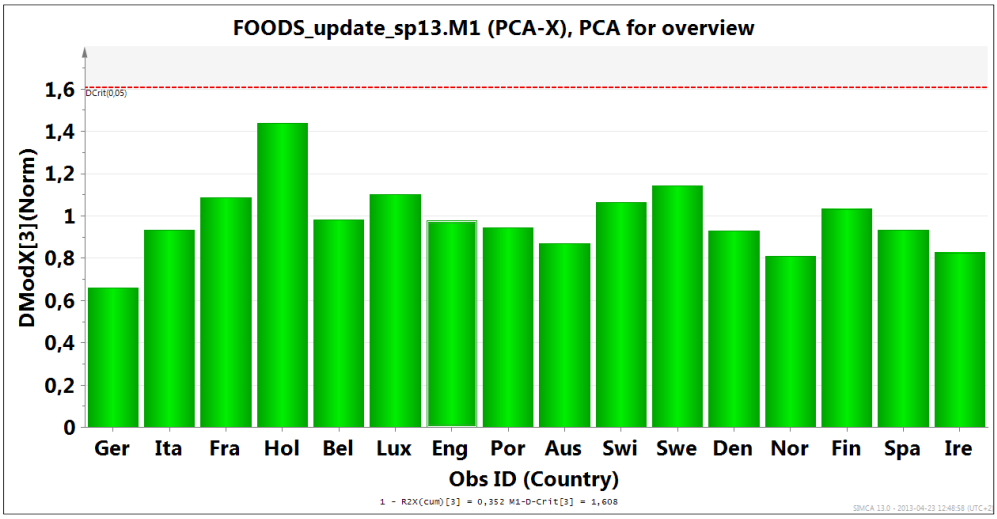


Figure 1.6: DModX plot of FOODS model. There are no outliers (all countries fit the model well).

The conclusion from analyzing the FOODS data was that strong characteristic consumption profiles are expressed in the data, and that these trends to some extent coincide with geographic location. This kind of information is hard to obtain by visual inspection of the raw data in Table 1.1. We will consider this example in more detail in Chapter 3.

1.5.2 A quality control dataset (CUPRUM)

Another application which we shall study for data overview purposes is called CUPRUM. An electrolysis process at Boliden AB (Skelleftehamn, Sweden) was sampled during one year of copper production by a continuous process. Two copper samples were taken each day, creating a data base with 730 samples (observations, rows). Each copper sample was analyzed for impurities, and the levels of eight metals (in ppm), viz., Ag, Ni, Pb, Bi, Sb, As, Te, and Se, were recorded (variables, columns). A scale called TAI had been developed to define the copper quality. TAI is a weighted sum of the eight different impurities; copper samples with TAI > 8 were considered bad and discarded. An excerpt from the dataset is given in Table 1.2.

Table 1.2: An excerpt of the CUPRUM dataset containing 730 observations and eight variables.

Dataset - Cuprum

	1	2	3	4	5	6	7	8	9	10	11
1	Primary ID	OBSNAM	kAg	kNi	kPb	kBi	kSb	kAs	kTe	kSe	TAI
28	27	01141A02109	9	0,7	0,5	0,1	1,6	0,5	0,1	0,6	3,27333
29	28	01141B12105	7	0,5	0,4	0,3	2,8	1	0,1	0,1	4,38333
30	29	01151A04213	9	1,3	0,8	0,3	2,5	1,5	0,1	0,8	5,67667
31	30	01151B14107	9	0,7	0,5	0,25	2,3	0,9	0,1	0,5	4,40667
32	31	01161A06211	6	0,2	0,6	0,1	2,5	1,2	0,1	0,1	3,64
33	32	01161B16211	6	0,9	0,6	0,25	3	1,3	0,1	0,1	4,68
34	33	01171A08107	8	0,3	0,7	0,3	3,1	1,5	0,1	0,3	5,27667
35	34	01171B18106	7	0,4	0,6	0,15	2,3	1	0,1	0,5	4,03
36	35	01181A10110	6	0,7	0,6	0,35	4	1,5	0,1	1	6,67333
37	36	01181B20214	7	2,3	0,5	0,4	3,9	1,6	0,1	0,2	6,36
38	37	01191B19213	7	0,5	0,9	0,2	1,8	0,8	0,3	0,1	3,7

The first objective of the data analysis was to explore whether multivariate methods could be used for quality monitoring. Figures 1.7 and 1.8 provide the primary results from the PCA of the 730 x 8 data table.

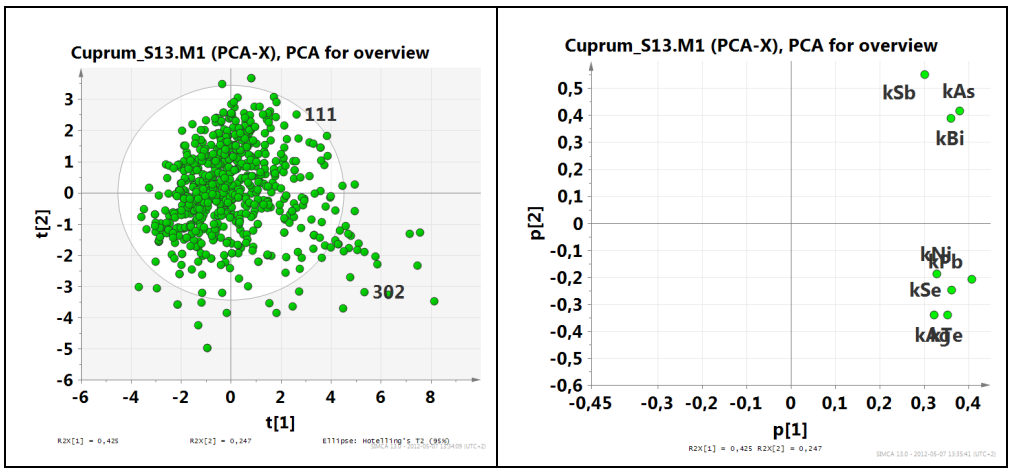


Figure 1.7: (left) PCA score plot of the CUPRUM dataset. Samples 111 (TAI = 8.1) and 302 (TAI = 7.8) are marked for reference purposes. Figure 1.8: (right) PCA loading plot of the CUPRUM dataset. Two groups of impurity variables are seen.

In Figure 1.7 samples 111 and 302 have been highlighted. These samples have similar quality according to the TAI scale. Sample 111 has TAI = 8.1 and will therefore be discarded, while sample 302 has TAI = 7.8 and will be sold as is. Interestingly, according to the PCA model these samples are modeled to exhibit totally different quality characteristics. This is because they are positioned in different parts of the score plot (Figure 1.7).

In order to interpret the PCA model, we may look at the loadings shown in Figure 1.8. Since all variables are found in the right-hand part of the loading plot, we understand that copper samples located in this area in the corresponding score plot have the lowest quality (they have the highest values of the impurity variables). Thus, the interpretation is that the horizontal direction represents average impurity of copper samples, and is able to discriminate between high quality (left-hand side) and low quality (right-hand side) products. The vertical direction, on the other hand, reflects deviations from the average impurity profile. Thus, there are two main impurity profiles, one where the copper samples have elevated levels of Sb, Bi and As (like sample 111), and the other where the samples have comparatively high levels of Ni, Pb, Se, Ag, and Te (like sample 302).

The conclusion of this study was that PCA enabled a better understanding of the quality characteristics of the copper produced. It was realized that the use of the single TAI scale was inefficient, even misleading, as this index did not appropriately distinguish between samples with different impurity patterns. PCA found these differences by considering all variables simultaneously. This example is further addressed in Chapters 3 and 7, where DModX and other important model parameters are also considered.

1.6 Illustration of problem types – Classification & Discrimination

In many datasets distinct clusters of observations (and variables) may be encountered. When there are many observations in each cluster, it might be difficult to analyze the fine detail within a cluster by means of the overall PCA model. In these cases local PCA modeling of each cluster can give better resolution. This brings us to the second main data analytical objective: classification and discrimination.

1.6.1 A classical dataset in statistics (IRIS)

One dataset which we will investigate in this context is the famous IRIS dataset, introduced by R.A. Fisher [Fisher, 1936]. This dataset comprises the lengths and widths of petals and sepals of three species of Iris: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. For each species, 50 observations (flowers) were studied, giving a complete dataset of 150 observations and four variables. A small selection is shown in Table 1.3.

Table 1.3: Part of the IRIS data with 150 observations and four variables.

Dataset - IRIS training

	1	2	3	4	5	6
1	Primary ID	OBSNAM	Se_Le	Se_Wi	Pe_Le	Pe_Wi
2	1	Setosa___1E	5,1	3,5	1,4	0,2
3	2	Setosa___1E	4,9	3	1,4	0,2
4	3	Setosa___1E	4,7	3,2	1,3	0,2
5	4	Setosa___1E	4,6	3,1	1,5	0,2
6	5	Setosa___1E	5	3,6	1,4	0,2
7	6	Setosa___1E	5,4	3,9	1,7	0,4
8	7	Setosa___1E	4,6	3,4	1,4	0,3
9	8	Setosa___1E	5	3,4	1,5	0,2
10	9	Setosa___1E	4,4	2,9	1,4	0,2
11	10	Setosa___1E	4,9	3,1	1,5	0,1

We applied PCA to these data in order to examine the overall data structure. Although classification might be the final objective, we recommend that PCA be applied to the complete dataset as a first step. According to the score plot seen in Figure 1.9, the species Iris setosa is quite different from the other two. This difference primarily arises because the petals of the Iris setosas are both shorter and slimmer than the others (see loadings in Figure 1.10). The sepals are also shorter. The distribution of the observations in the vertical direction of Figure 1.9 primarily reflects the changing widths of the sepals. Wide sepal observations are found in the upper part of the score plot, and slimmer specimens in the lower part.

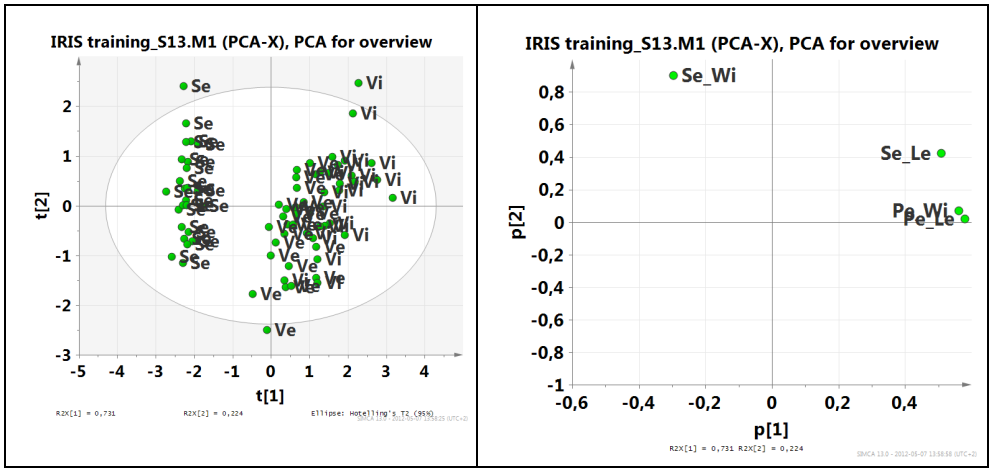


Figure 1.9: (left) PCA score plot of IRIS dataset. Figure 1.10: (right) PCA loading plot of IRIS dataset.

In classification one approach is to establish local models for each class of observation (here: each species of Iris). With such local models it is then possible to assess the probable class identity of additional samples of Iris, not previously included in the data analysis. Here, it seems possible to diagnose what is Iris setosa and what is not. However, it is not easy to separate Iris versicolor from Iris virginica, because of a considerable overlap in the score plot (Figure 1.9). One way of addressing this problem is to make a new PCA model from which Iris setosa specimens are excluded. If the overlap still persists it might be necessary to treat the latter two species together. More detailed results from these classification efforts are given in Chapter 10.

1.6.2 A second dataset for classification and discrimination (ARCHAE)

The second main dataset for classification and discrimination (ARCHAE) deals with archaeology data. Johan Linderholm at Umeå University collected 22 soil samples at an excavation site outside Mjölby, Sweden, and characterized them with 18 variables reflecting trace element and main component composition [Linderholm, *et al.*, 1994]. The collected samples were of three different origins (Figure 1.11), denoted *Feature* when coming from recognizable features (refuse pit or hearth) of the dwelling site, *Site* when originating from the occupational layer on the site, and *Control* when originating from an off-site control sample. The objective of the study was to develop a method of classifying archaeologically related soil samples by means of multivariate analysis of soil composition data.

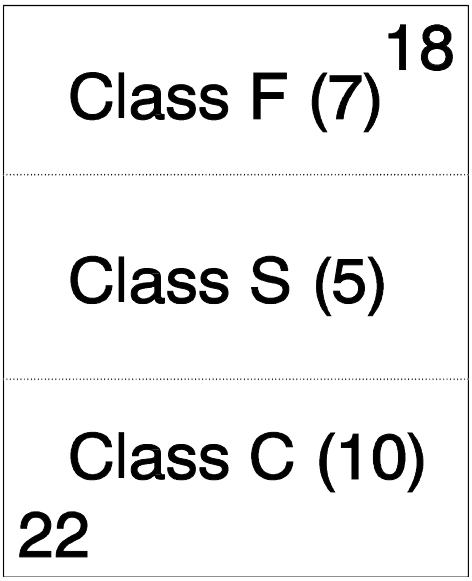


Figure 1.11: The 22 archaeological soil samples represent three categories of soil: feature (F), site (S), and off-site controls (C). The data analysis will reveal whether the collected data convey any class separating information.

With PCA it is easy to see whether the data contain class separating information. Figures 1.12 and 1.13 display the score and loading plots obtained by analyzing the data.

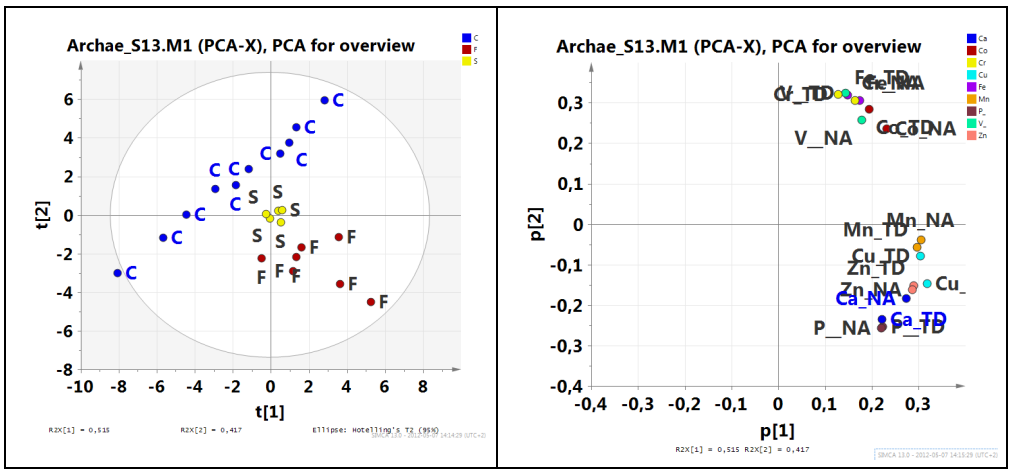


Figure 1.12: (left) PCA score plot of the ARCHAE example. Each soil sample is a plot mark. The plot labels C, S, and F denote control, site, and feature, respectively. We can see that the three classes of soil sample are separated. This suggests that the soil composition data carry class separating information. Figure 1.13: (right) PCA loading plot of the ARCHAE example showing two distinct patterns of trace element and main component profiles. Evidently, trace elements like Mn, Cu, and Zn covary with the main component phosphate (P). The other conglomerate of variables is made up of Co, V, Cr and Fe.

There is no doubt that the trace element and main component concentrations carry information that is related to the nature of the soil samples. Linderholm concluded that multivariate analysis opens possibilities for discriminating between undisturbed soil and soil modified due to cultivation and dwelling. We could then establish one PCA model for each type of soil, and make classification of the likely origin and history of future soil samples from archaeological excavations. We will return to this example in Chapter 10.

1.7 Illustration of problem types – Regression modeling

The ability to predict response variables (Y) from an assembly of predictor variables (X) is perhaps the most intriguing and important objective in multivariate data analysis. Regression analysis based on multivariate techniques is common in design of experiments (DOE) applications, multivariate calibration, process monitoring, structure-property modeling, and so on. In this book we will provide ample documentation of such applications, all using PLS regression (or extensions thereof) as the data analytical tool.

1.7.1 A multiple response optimization dataset (LOWARP)

The first PLS example is called LOWARP. This example concerns the development of a polymer similar to that used in the plastic covering of mobile phones. The desired profile of the polymer was low warp and high strength. Four constituents (glas, crtp, mica, and amtp) were varied in the polymer formulation by means of a 17 run mixture design. For each new polymer, i.e., each new experiment in the mixture design, 14 responses relating to both warp and strength were measured on the product, as shown in Table 1.4. The objective of the data analysis was to uncover which combination of factors (the four ingredients) gave polymers with low warp and high strength.

Table 1.4: The LOWARP dataset. The first four columns are the factors. Variables 5-18 represent the 14 responses. Each row of the data table represents one polymer.

		1	2	3	4	5	6	7	8	9	10	13	16	11	12	14	15	17	18
Num	Name	glas	crtp	mica	amtp	wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	wrp7	wrp8	st1	st2	st3	st4	st5	st6
1	1	40	10	10	40	0.9	5	0.2	1	0.3	4.2	1.2	1.3	232	15120	2190	26390	2400	0.7
2	2	20	20	0	60	3.7	7.3	0.7	1.8	2.5	5.4	1.8	2.1	150	12230	905	20270	1020	0.6
3	3	40	20	0	40	3.6	6.9	0.9	2.1	4.8	9.4	1.2	1.4	243	15550	1740	21180	1640	
4	4	20	20	20	40	0.6	3.1	0.3	0.4	0.4	1.1	1	1	188	11080	1700	17630	1860	0.5
5	5	20	10	20	50	0.3	2.1	0.3	0.3	0.8	1.1	1.2	1.3	172	11960	1810	21070	1970	0.5
6	6	40	0	20	40	1.2	5					1.1	1.3	245	15600	2590	25310	2490	0.6
7	7	20	0	20	60	2.3	3.9	0.3	0.4	0.7	1.4	1.5	1.6	242	13900	1890	21370	1780	
8	8	40	0	10	50	2.6	5.9	0.4	0.2	0.7	1.2	1.6	1.6	243	17290	2130	30530	2320	0.7
9	9	30	20	10	40	2.2	5.3	0.2	0.7	0.6	2	1	1.1	204	11170	1670	19070	1890	0.6
10	10	40	0	0	60	5.8	7	0.9	1	5.6	11.8	1.6	1.8	262	20160	1930	29830	1890	
11	11	30	0	20	50	0.8	2.9	0.5	0.6	1.1	2	1.3	1.3	225	14140	2140	22850	2110	0.7
12	12	30	10	0	60	2.8	5.1	1	1.2	2.7	6.1	1.9	2.1	184	15170	1230	23400	1250	0.6
13	13	30	10	10	50	1.1	4.7	0.6	0.9	1.3	3.5	1.4	1.4	198	13420	1750	23790	1930	0.7
14	14	30	10	10	50	1.9	4.7	1	1	2.8	5.4	1.5	1.6	234	16970	1920	25010	1790	0.7
15	15	30	10	10	50	2.9	5.9	0.5	0.6	1	6.6	1.5	1.6	239	15480	1800	23140	1730	
16	16	40	10	0	50	5.5	7.9	0.8	2.4	5.5	9.3	1.5	1.8	256	18870	1880	28440	1790	
17	17	30	0	10	60	3.2	6	0.3	0.5	1.5	5.2	1.5	1.7	249	16310	1860	24710	1780	

PLS was applied to the LOWARP data. One powerful feature of PLS is that it allows many responses to be incorporated in one regression model. This is possible so long as the responses are reasonably well correlated and thus give similar information about the observations. As with PCA, PLS also produces sets of score and loading ("weight") plots. One informative weight plot is given in Figure 1.14, which shows the relationships between the four X-variables and the 14 Y-variables at the same time. Note that the interpretation of a PLS weight plot is outlined in Chapter 4.

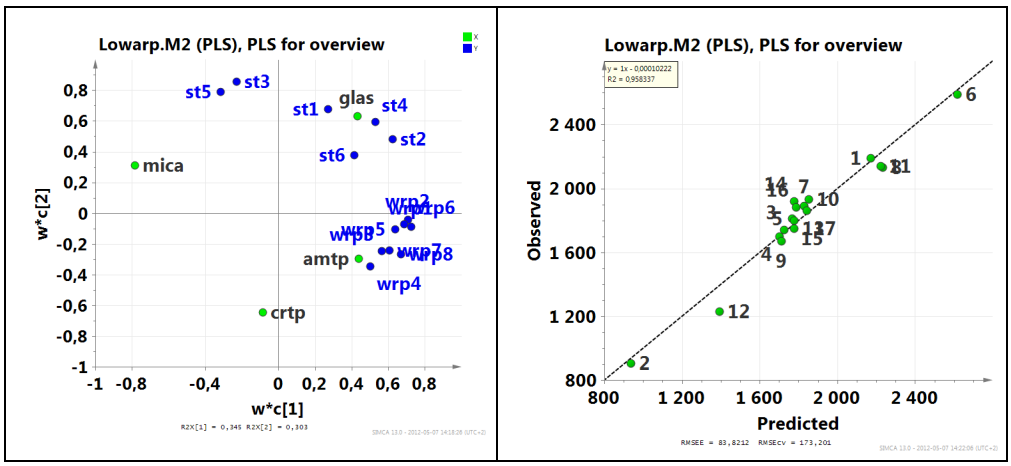


Figure 1.14: (left) PLS weight plot of the LOWARP example. This plot reveals the inter-relatedness between the four factors (mica, amtp, crtp and glas) and the 14 responses. Figure 1.15: (right) Observed versus predicted values for the response st3. Each point denotes one polymer of the 17 run mixture design.

As seen in Figure 1.14, the eight warp responses are situated tightly together. This indicates that they are strongly correlated. The six strength responses, however, are less well correlated. Two strength responses, st3 and st5, have little correlation with the warp responses, whereas the other four (st1, st2, st4 and st6) are partially correlated with the warp responses. We can see that varying the glas factor is unlikely to help us reach our goal, as it affects both strength and warp in the same way. Increasing the amount of glas will increase the values of all responses. In order to increase strength we should focus primarily on lowering the amount of crtp in the recipe, as this factor does not influence warp to any appreciable extent. To decrease warp we should increase mica and decrease amtp.

PLS provides many plots and parameters which can be considered when assessing the quality of a model. One popular plot is the observed/predicted plot for each individual response. Figure 1.15 shows this relationship for the response st3. Apparently, there is good agreement between the measured data and the estimates from the derived PLS model.

The final conclusion of this investigation was that the four factors are highly informative for explaining the responses. More information is given in Chapter 4.

1.7.2 A multivariate calibration dataset (SUGAR)

Our second regression case study involves the use of PLS for multivariate calibration. We call this application SUGAR, as it concerns fluorescence measurements from the industrial production of white sugar. This example was kindly contributed by Rasmus Bro at KVL in Copenhagen [Bro, 1996]. This dataset well exemplifies how any spectroscopic curve can be transferred to a single point in a score plot, an operation which will reveal the features of the spectrum in relation to all other spectra that are analyzed.

At a sugar plant somewhere in Scandinavia an important quality parameter was the ash content of the sugar, but its measurement was time-consuming. Consequently, it was desired to replace the conventional wet-chemistry measurements with rapid on-line fluorescence measurements. To investigate whether this was feasible multivariate calibration, linking the spectral data (X) to the wet-chemistry measurements (Y), was attempted. The collected spectral data are plotted in Figure 1.16.

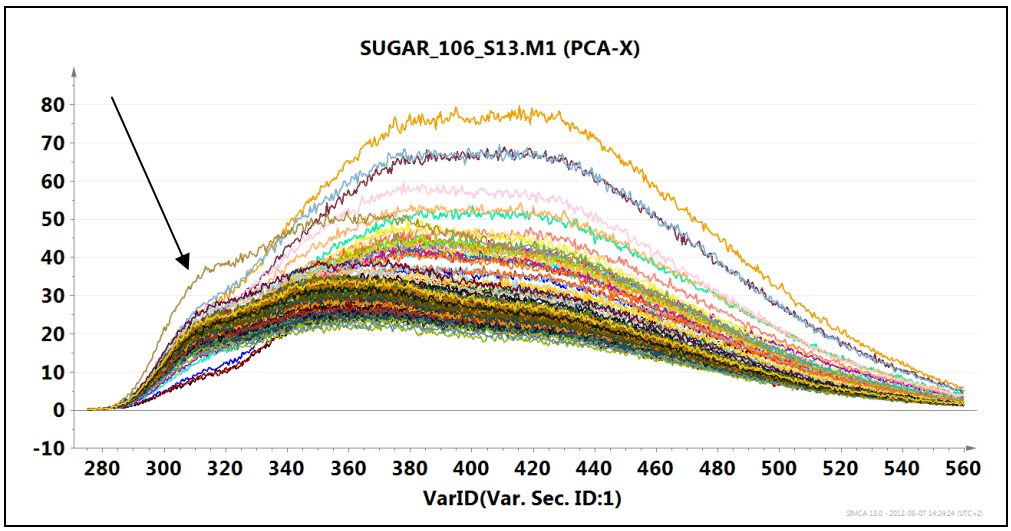


Figure 1.16: Plot of raw fluorescence data. Each spectrum, when digitized, corresponds to one row of the X-matrix. The arrow indicates an atypical spectral behavior of sample number 88. The spectral range is 275-560nm.

The data were measured during the very first 2.5-day campaign at a new factory. A total of 106 samples were examined and by digitizing the fluorescence spectra 571 X-variables were compiled for each.

Initially, only the spectral data were investigated with PCA. This revealed a distinct shift in the process expressed in the data (Figure 1.17). The score plot of Figure 1.17 indicates that when the new process started it was unstable. After a while an area with more stable and chemically meaningful operating conditions was reached. This is also the area for which calibration modeling is warranted.

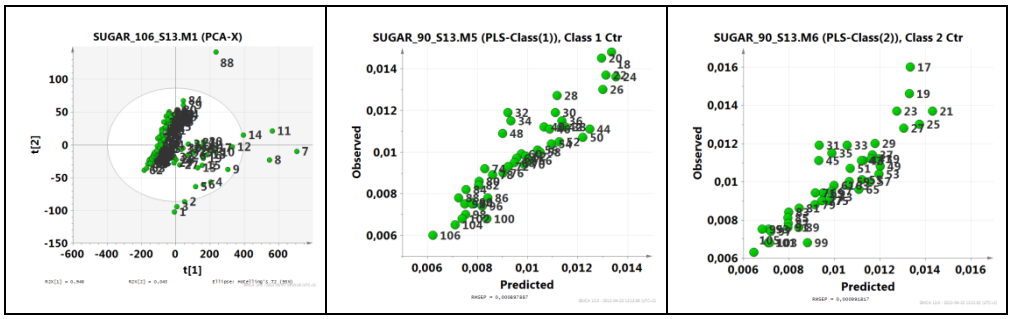


Figure 1.17: (left) PCA score plot of the SUGAR dataset. Each point represents one observation (time point). Apparently, a strong process shift occurs at the beginning of the sampling period. Sixteen samples were judged to be uninteresting and were therefore omitted from further modeling. Figure 1.18: (middle) Relationship between observed and predicted ash content when model training was made with odd-numbered samples and model validation with even-numbered samples. Figure 1.19: (right) Same as Figure 1.18, but the roles of the data sub-sets have been reversed.

Accordingly, the first 16 samples (time points) were discarded and PLS was then used to relate the spectral data to the ash content. In order not to overfit the calibration model, the data were divided into two parts, odd and even-numbered samples. Subsequently, one part of the data was used for model training and the other part was used for model testing (validation), and vice versa. Figures 1.18 and 1.19 give the prediction results. As seen, the two PLS models predict the ash content well.

The conclusion was that the fluorescence data could model and predict the ash content. The calibration models obtained enabled on-line predictions of an important quality parameter reflecting the state of the process and the quality of the sugar.

1.8 Questions for Chapter 1

1. Why is multivariate data analysis useful?
2. What is an observation?
3. What is a variable?
4. What is indicated with the acronym PARC?
5. Can you list three basic data analytical problem types?
6. Which multivariate data analytical tool is useful for data overview?
7. Which multivariate data analytical tool is useful for regression modeling?

1.9 Summary and discussion

In this chapter we have provided an introduction to multivariate data analysis (MVDA). MVDA is becoming increasingly popular because the on-going data explosion tends to overload our computers and data-bases with massive amounts of data. In MVDA, two data analytical tools, PCA and PLS, are particularly important. One nice feature of these methods is that the presentation of the modeling results is graphically oriented, and that many diagnostics and parameters are available for model interpretation and validation.

PCA is useful for overviewing a data table, because it is able to explore relationships both among variables and observations. PLS also has this ability, but is more useful when two sets of variables, X and Y, have to be connected with the aim of predicting Y for new observations based on known X-data. PCA and PLS make it possible to address three main types of data analytical issues, i.e., (i) overview of data, (ii) classification and discrimination, and (iii) regression modeling.

2 Basic concepts and principles of projections

2.1 Objective

The objectives of this chapter are (i) to describe some basic concepts used in multivariate data analysis, and (ii) to introduce the principles behind projection methods. To this end, we shall consider the concepts of *data* and *variability*, and their consequences for multivariate data analysis. Three approaches to "analyzing" data tables, the COST approach (consider one separate variable at a time), classical statistics, and multivariate data analysis, are discussed. The *principles of projections* are outlined towards the end of the chapter, and in connection with this discussion the *concept of a model* and the *philosophy of model building* (modeling) are considered. The FOODS dataset will be utilized as an illustration.

2.2 Data

2.2.1 A table of data is not the same as information

In order to understand the essential features of a system or a process it is necessary to make measurements. Such measurements might be temperatures inside a reactor, material flows from a mine, or NIR spectroscopic characterization of the coating layer of a tablet. These measurements are called *data* and are collected to understand the properties of the modeled system or process. It is important to realize that the data contain both systematic information about the system and unwanted variation (noise). One way to formulate this is

$$y = \eta + \varepsilon \qquad \text{(eqn. 2.1)}$$

where y is the data which is modeled by η, the deterministic component, and ε, the stochastic component. The former represents a function [f(x)] approximating y, and the latter comprises the noise. The noise is a mixture of measurement errors, experimental errors, sampling errors, model errors and other sources of variability. The mere fact that we have made a lot of measurements does not *per se* imply that we know the properties of a system. A data table must be analyzed by appropriate methods to reveal the systematic information.

2.2.2 Variability and types of data

There are a few points which must be considered in connection with the concept of data: (i) variability, (ii) complexity of the data, and (iii) types of data. *Variability* arises in all applications, as all measurements and experiments are influenced by noise (see Figure 2.1). Of course, systematic variability can also be introduced into the data by actively changing some important factors, for instance according to a design of experiments (DOE) protocol. The presence of unwanted and uninformative variability in the data must always be considered, to avoid drawing misleading conclusions. Multivariate analytical methods are well suited to dealing with variability in large data tables, thereby reducing the risk of incorrect inferences.

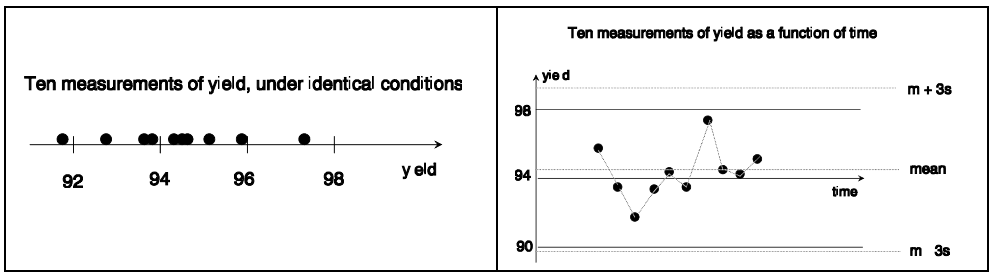


Figure 2.1: An illustration of variability. (left) Any measurement or experiment is influenced by noise. (right) Under stable conditions, any system or process varies around its mean, and stays within "control limits", usually ± 3 standard deviations.

The second point concerns the *complexity* of the data. A data table starts to become incomprehensible to a superficial overview once the number of variables exceeds about five; at this point it becomes "multivariate". As long as the number of variables does not exceed five, the information in a data table can be evaluated by making histograms and calculating averages, variances, covariances, and so on. With larger datasets these simpler statistics become more difficult to overview. Figure 2.2 lists typical complexities of data.

Complexity of data	**Number of variables, K**
Univariate	$K = 1$
Bivariate	$K = 2$
Low variate	$K \leq 5$
Multivariate	$K \geq 6$

Figure 2.2: Classification of data according to number of variables, K.

Another way to categorize data is based on the nature of the variables. In a broad sense, variables may be divided into *factors* and *responses* (Figure 2.3). The *factors*, sometimes also known as predictors or parameters, are variables whose different levels *might* exert an influence on the system or the process. In regression, these form the X-matrix. The *responses* are the variables which are measured to capture the performance of the system. Examples are the amount and quality of a product, energy consumption, production costs, or level of pollution. In regression, these form the Y-matrix.

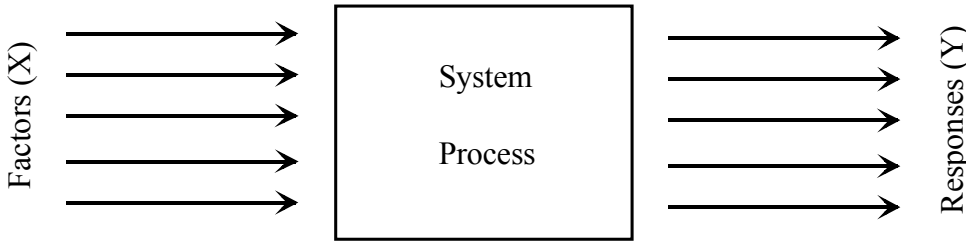


Figure 2.3: Variables can be classified as factors (X) or responses (Y). The measured responses describe the properties of the investigated system. The variation in X may be related to Y by multivariate data analysis.

Furthermore, variables (factors and responses) are either *quantitative* or *qualitative*. A quantitative variable may assume any (reasonable) numerical value along a continuous scale. Therefore, such variables are also called *continuous* variables. On the other hand, a qualitative variable is a categorical variable, which can only assume certain defined levels. A typical qualitative factor might be the type of disintegrant used in making a tablet. For instance, such a factor might assume the three levels: disintegrant A, disintegrant B, and disintegrant C.

Finally, in DOE, a division into *controlled* and *uncontrolled* factors is common. Controlled factors can be set to a desired value and be kept there. Uncontrolled factors are those which are difficult to regulate, but which may have an impact on the result. Typical examples are

ambient temperature and humidity. We strongly recommend keeping track of such uncontrolled factors.

2.3 Some properties and challenges of multivariate data

There is a steady trend in the modeling of chemical and other manufacturing processes to use more and more variables to characterize the molecules, reactions, processes, samples, or other "systems". By means of sensor arrays, data characterizing the state of the process are collected at regular time intervals (ranging from several hundred hertz for some steps in semiconductor industry, to once per minute for fairly fast processes such as distillations, and to once every six hours for very slow fermentations). In addition, samples are occasionally withdrawn and analyzed for product yield, purity, side products, etc.

There are many reasons for this trend. Firstly, we instinctively feel that we know and understand more about our systems when we have measured many properties (variables) rather than few. Secondly, electronics, computers, and the "instrumental revolution" provide the opportunities to obtain a lot of data from any investigated system. However, this abundance of data leads to greater analytical difficulties than those associated with a more limited number of variables.

2.3.1 All data are needed

The first and most important point is that *all data points are needed*. One should not discard data, or analyze variables one by one, or select just a sub-set of the “best” variables. The reason is that variables are often collinear, either partially or completely, and that the information is found in the correlation pattern rather than in the individual signals (see Figure 2.4). If parts of the data, for some reason, are ignored, there is a substantial risk of overlooking important information.

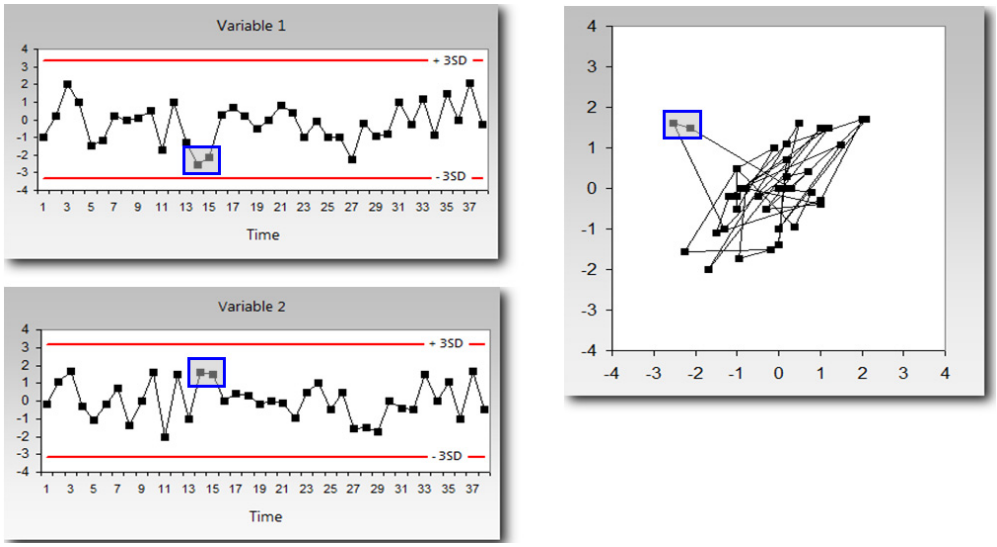


Figure 2.4: In the early days of process analysis (1920s and 1930s) relatively few measurements of product quality were made. Such quality measurements were often tracked over time in control charts like the Shewhart plot. In the example, two variables are charted over time, and for each variable two measurements have been marked with solid squares. According to the individual variables, these two samples are good in every aspect since they are found well inside the control limits, located at ± 3 standard deviations. However, the two highlighted samples appear as outliers when the two variables are plotted against each other. Most outliers are NOT detected until all variables are analyzed together. This is the key to early fault detection.

2.3.2 Other features

Some of the other challenges and difficulties with complex and multivariate data are the following:

- *Dimensionality*
  In process supervision and status visualization, more than a thousand variables might be registered every few seconds, and 10-20 quality variables every few minutes. This "data overload" is difficult to cope with using traditional analytical approaches; often the operators concentrate on only a few reference variables and ignore the remaining ones. Multivariate data analysis seeks to overcome this limitation by using all variables at the same time. An important point is that multivariate projection methods work with long and lean ($N >> K$), short and fat ($N << K$) and almost square ($N \sim K$) matrices.
- *Multicollinearity*
  Although thousands of variables might be available, there are certainly not a thousand independent events taking place in the system or process. Usually, only a few underlying phenomena affect the system, and the question is which ones (see also Section 2.4). The interpretation of relationships is complicated by the multiple collinearities that often occur among variables. Multicollinearity arises because some variables are approximate linear functions of other variables. This appears automatically in short and wide matrices (few observations, many variables) even though the pair-wise variable correlations themselves may be low or moderate. The consequence of multicollinearity is that projection methods are needed to treat the data appropriately. It is not meaningful to use multiple linear regression (MLR) or similar methods, since they require that the rank of the data is K ("full rank").
- *Noise*
  Variation introduced by disturbing factors, both known and unknown, is called experimental error, or simply noise. Because individual measurements may be noisy and contain large variability, important effects may be completely or partially obscured. Fortunately, multivariate projection methods are robust to noise, and enable a filtering of the data such that the systematic information (the "effects") is highlighted. This because the latent variables become stabilized by including many relevant variables, even if each of the measured variables is very noisy.
- *Missing data*
  Process data tables often have gaps, e.g., due to transducer problems or a failing sensor. Projection methods tolerate moderate amounts of missing data both in X and in Y. The larger the matrices X and Y, the higher the proportion of missing data which can be tolerated. For ordinary sizes of data with around 50-100 observations, around 10 to 20 % missing data can be handled, provided that they are not missing according to a systematic pattern [Wold, *et al.*, 1984].

2.4 Confusion of correlation and causation

It is necessary to distinguish between the concepts of *correlation* and *causation* [Box, *et al.* 1978]. Figure 2.5 shows a scatter plot of two variables -- whose identities will be explained shortly -- producing a typical pattern of data association. These two variables are positively correlated, which means that when the values of the first variable are raised, the values of the second variable increase too. Variables may also be negatively correlated, indicating that when one variable increases the other decreases. In multivariate data, the possibility of finding such correlations by chance increases dramatically with the increasing number of variables. Hence, suitable analytical methods, able to deal with correlated variables, should be used.

Causation is fundamentally different from correlation. A causal relationship implies that a change in one variable, a factor, *caused* another variable, a response, to increase or decrease. Causal relationships are best established through the use of design of experiments. Such relationships are often hidden in multivariate applications because of the sheer amount of data.

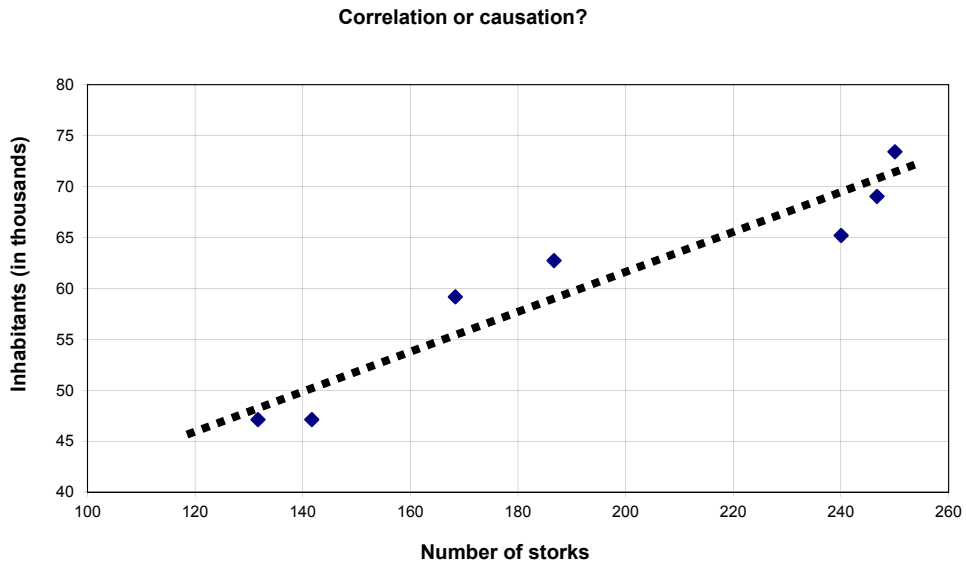


Figure 2.5: A scatter plot of the number of inhabitants of the German city of Oldenburg in the years 1930-1936 versus the number of storks observed each year in that period [Box, et al., 1978].

Now, we turn to Figure 2.5. As shown by this scatter plot, there is a strong association between the number of inhabitants of the German city of Oldenburg in the years 1930-1936 and the number of storks observed each year in that period. Here, a causal relationship would mean that it was actually the increased number of storks that caused the population of Oldenburg to increase. Although in this example few analysts would infer this causal relationship, investigators frequently make this kind of misinterpretation in other contexts. In fact, the correlation between the number of storks and the size of the population arises because these two variables are both linked to a third, unknown factor, which also happened to change during the 7-year period.

2.5 How do we analyze our data today?

Any data table must be analyzed to extract the information in the data. We divide the methods used into three levels of sophistication and ambition. These are (i) the COST-approach, (ii) classical statistics, and (iii) multivariate analysis by projection methods.

2.5.1 The COST approach

The intuitive way of looking at data tables is to consider one separate variable at a time (COST). The COST approach is sometimes also called OVAT (one-variable-at-a-time) or “eye-balling”. For instance, to distinguish between two classes of observation, we start with one variable, compute average and standard deviation for each class, and then test for the class separation in that variable. This is followed by an investigation of variable two, then variable three, etc..., in the same manner. Usually, none of the examined variables seem to provide any information about the problem. Similarly, when investigating relationships between two sets of variables, X and Y, one looks at all pair-wise correlations X_k/Y_m.

Thus, the COST approach is an inefficient method with a serious drawback: the information in multivariate data often remains hidden. There is a second and even greater problem with COST, that of spurious results. The "significance" of the difference in a variable between two classes is usually evaluated at the 5 % probability level. Thus, a difference between the two classes is judged to be “significant” when there is less than 5 % risk that this difference may be due to chance alone. But when a difference is found in one of two variables this risk is about twice as large; and with one of K variables the risk rises to $1-(0.95)^K$. This risk rate is outlined in Figure 2.6.

- Risk of spurious results (Type I errors, false positives)

  Risk = $1 - 0.95^K$

K	5	10	20	40	60	80	100
Risk	0.23	0.40	0.64	0.87	0.95	0.98	0.99

- Risk of not seeing the information (Type II errors, false negatives)

Figure 2.6: Overview of the problems of the COST-approach. K represents the number of variables of a data table.

As seen, for K = 20 (as in the FOODS example), the risk is 64 % and for K = 60 it is 95 %. Therefore it is very difficult to judge whether the differences in single variables between classes are real or not. The only way to handle these two problems, also called Type I and Type II errors, is to analyze all variables simultaneously.

2.5.2 Classical methods of statistics

The next level of sophistication comprises the classical methods of statistics. An example is multiple linear regression, MLR, a commonly used regression tool. MLR, canonical correlation, analysis of variance, and others, were developed for the type of data matrices of the early days of the 20th century. At that time it was difficult to make extensive measurements on a series of samples. Thus, data tables usually had many more observations than variables. The IRIS dataset is a good example. We refer to such matrices as being "long and lean" (Figure 2.7).

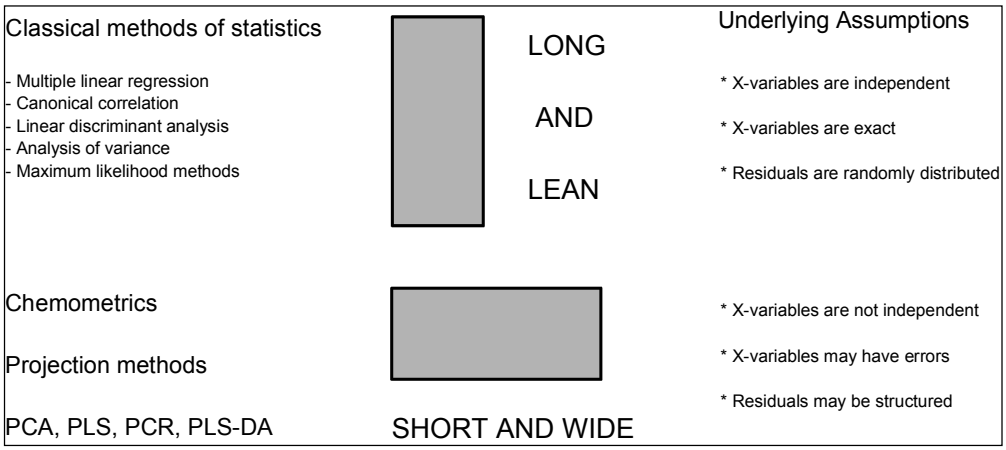


Figure 2.7: Two shapes of data table, long and lean, and short and wide, and some assumptions of common analytical methods used to deal with these matrices.

The classical methods assume statistical independence of the X-variables, and that the X-variables are exact and completely (100%) relevant. The first assumption is met reasonably well by long and lean matrix structures, but is violated as soon as the number of variables approaches or exceeds the number of observations. MLR also has problems with regard to missing data. As soon as some data element is lacking that row and that column must be eliminated from the modeling.

However, MLR is an excellent technique for data analysis in design of experiments (DOE) applications, particularly with few responses.

2.5.3 Multivariate data analysis

Today, the reality for experimentalists has changed. Data matrices are no longer long and lean, but rather short and wide (Figure 2.7). This causes problems for the classical methods of statistics and raises new demands on the data analytical techniques. Short and wide data structures arise because it is no longer difficult or time-consuming to measure variables.

Modern electronics has allowed a vast array of instruments (spectrometers, chromatographs, etc.) to be developed, which are capable of outputting hundreds or thousands of variables in a short time. The number of observations, on the other hand, is often more difficult to increase. This is because (new) regulations apply in certain fields regarding costs, time and ethics (individuals, animal testing, etc.), which constrain the number of samples.

Large data tables in general, and short and wide matrices in particular, require the use of multivariate projection methods, like PCA and PLS. Multivariate analysis by projections has the following advantages:

- Deals with the dimensionality problem
- Handles many variables and few observations (short and wide data tables)
- Handles few variables and many observations (long and lean data tables)
- Handles almost square matrices of any size
- Copes with multicollinearity
- Copes with missing data
- Is robust to noise in both X and Y
- Separates regularities from noise
- Provides informative diagnostic and graphical tools

In summary, there exist projection methods for each of the three major data analytical problems (overview, classification, and regression) discussed in Chapter 1. Thus, multivariate projection methods are versatile and useful. It is important to recognize, however, that MVDA does not deal with how to structure a problem or how to plan which experiments to perform; this is dealt with by design of experiments (DOE) techniques.

2.6 The principles of projections

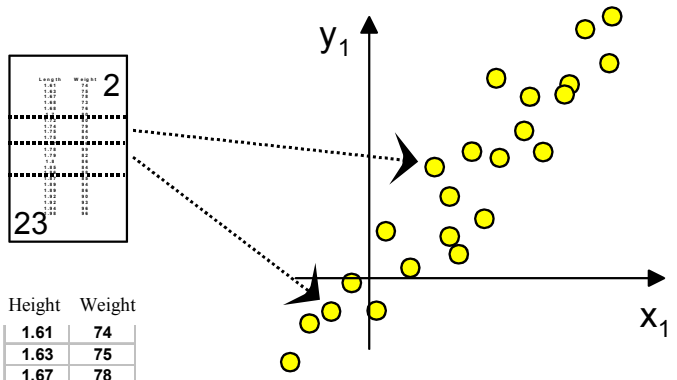
In this section, we will outline the principles of projections. Detailed accounts of the PCA and PLS methodologies are given in Chapters 3 and 4.

2.6.1 K-space and M-space

The basics of MVDA can be outlined in terms of the geometry of multidimensional spaces. Multivariate data and models can be represented as points, lines, planes, and hyperplanes in these spaces. We have a good ability to comprehend structures in two- and three-dimensions. The understanding of these few dimensions can then be generalized to higher dimensional spaces spanned by many variables. We call such a high-dimensional space the K-space if we are working with X-variables, and the M-space if Y-variables are considered. (Sometimes M-space also means “multidimensional space”.)

2.6.2 Example K = 2

The simplest situation is when we only have two variables ($K = 2$). Two variables, for instance x_1 and y_1, in linear regression analysis define the familiar x-y co-ordinate system (Figure 2.8). This co-ordinate system can be used to plot the data of two variables, as exemplified by Figure 2.8. In this example, the relationship between the body height and body weight of 23 individuals is examined. Evidently, there is a strong relationship between these two variables, suggesting that when the body height increases the body weight is raised as well (and vice versa).



Height	Weight
1.61	74
1.63	75
1.67	78
1.68	73
1.68	76
1.7	80
1.72	80
1.74	79
1.75	84
1.75	80
1.77	86
1.78	99
1.79	82
1.8	86
1.85	84
1.86	88
1.87	85
1.89	94
1.89	86
1.92	90
1.92	93
1.94	96
1.98	96

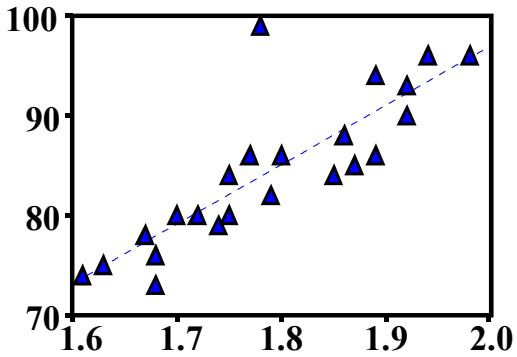


Figure 2.8: A two-dimensional co-ordinate system displays the data of two variables. In the lower part of the figure, we show the relationship between the body height and the body weight of 23 individuals. Evidently, there is a strong relationship between these two variables.

2.6.3 Example K = 3

Assume that we have a data table **X** with $K = 3$ variables denoted x_1, x_2, and x_3. Each X-variable then defines a co-ordinate axis in a three-dimensional co-ordinate system (Figure 2.9). Analogously, with K variables we would get a K-dimensional co-ordinate system.

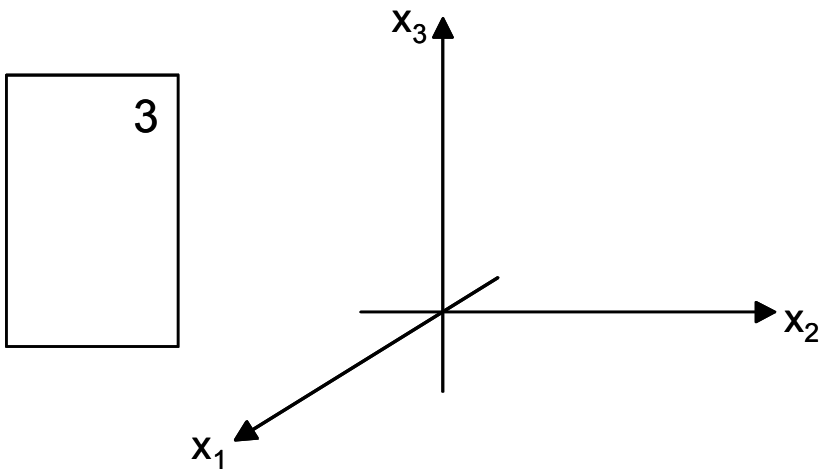


Figure 2.9: A three-dimensional co-ordinate space spanned by the variables x_1, x_2, and x_3.

2.6.4 Each row of a data table is interpretable as a point in a multidimensional space

Each observation in **X** is then characterized by three numerical values. The three values of an observation are represented by one point in this space. This is illustrated in Figure 2.10.

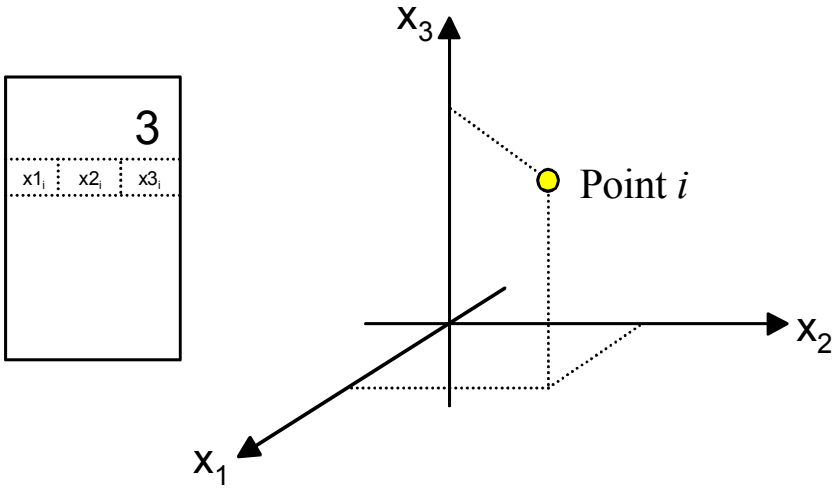


Figure 2.10: Each row in a data table defines a point in space spanned by the K variables (here: K = 3).

Three variables represented by a three-dimensional space can be extended to *K* variables in a K-space. A K-space is constructed by letting each of the *K* X-variables define one orthogonal axis in a co-ordinate system. Although such a space cannot be easily visualized, it has a real existence analogous to that of the two- and three-dimensional spaces, and the properties of points, lines, planes, distances, etc., are the same in K-space as in 2-space or 3-space. Hence, we can use the latter as "models" of the former and illustrate multivariate analysis in terms of less complex spaces.

2.6.5 A data table forms a point-swarm

The principle of representing rows as points in space makes it possible to convert a data table to a graphical representation. All the observations of **X** are displayed in K-space as a swarm of points. For instance, an X-matrix consisting of 20 observations (rows) corresponds to a swarm of 20 points (Figure 2.11).

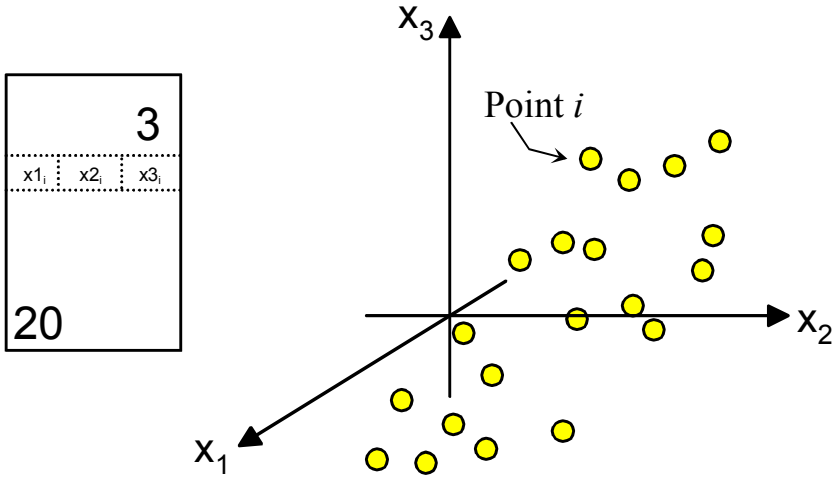


Figure 2.11: The data table forms a swarm of N points in K-space, here 20 points in a three-dimensional space.

2.6.6 Constructing a window into the three-dimensional space

Now, analyzing the data would correspond to formulating a quantitative description of the shape of the point-swarm. This is possible by using a mathematical model which approximates the data. Such a model can be viewed as a window in the three-dimensional space (Figure 2.12). This window is oriented such that it provides a good overview of the data and hence enables their interpretation. In this way, it is possible to convert a data table into a plot, which is informative and easy to understand. The principle behind this conversion process is called *projection*. We have accomplished a projection from the original three-dimensional space down onto a two-dimensional space - a window. By examining the location of each observation within this window we get an overview of the data.

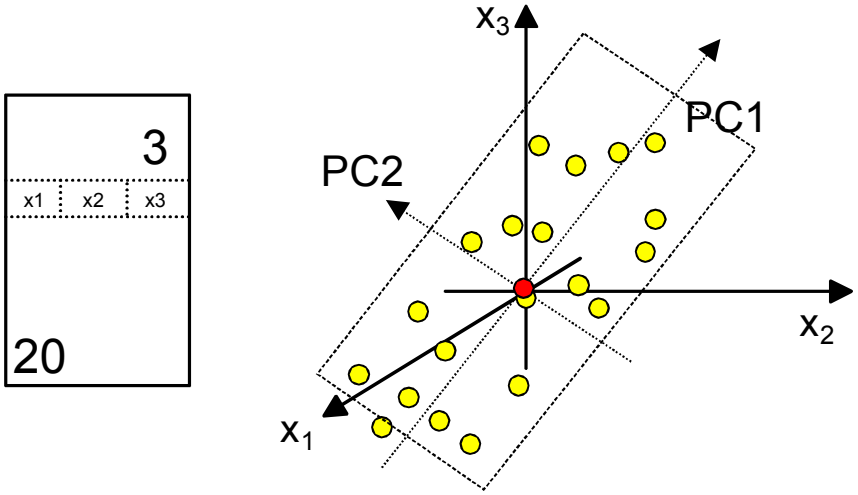


Figure 2.12: A window is inserted into the three-dimensional space. The observations (points) are projected onto this plane, which enables data overview and interpretation. This is the main step of principal component analysis, PCA.

2.6.7 Projections in higher-dimensional space

The idea of projections can now mathematically be extended from two and three dimensions to any number of dimensions, i.e., any number of variables (columns in the data table). Thus, a two-dimensional window (a plane) can be placed in the K-space and the data points projected onto this plane even when K is 7 or 33 or 4012. And this plane, this window, can be lifted out and displayed in a book page, or on a computer screen.

In the FOODS example, the K-space has 20 dimensions, because there are 20 variables (food items) in this dataset. Each of the 16 observations (nations) becomes one point in the twenty-dimensional space, and together they form a swarm of 16 points. The shape of the point-swarm can then be approximated by a low-dimensional model, a window (or plane), by projecting each observation onto that plane (Figure 2.13). This is the essence of PCA (principal component analysis).

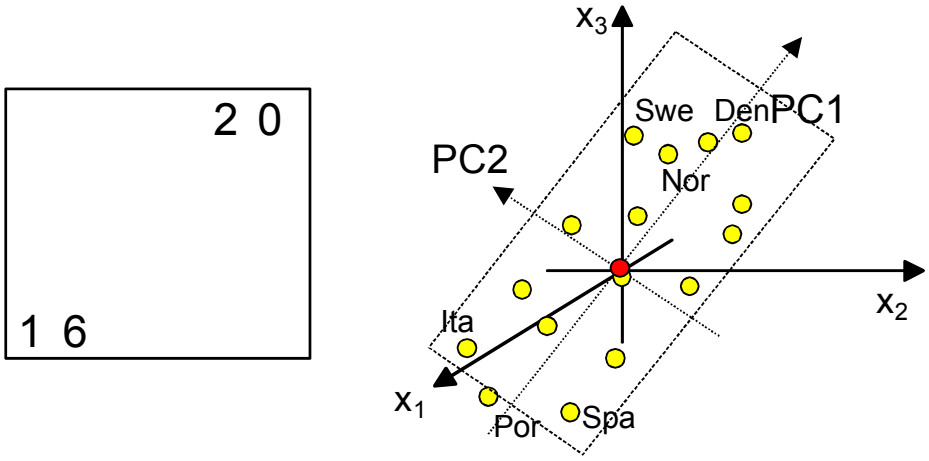


Figure 2.13: The 16 observations (countries) of the FOODS dataset form a swarm in a 20 dimensional space. For reasons of simplicity, the drawing is made to contain only three dimensions. By plotting the computed plane in terms of a scatter plot, it is possible to overview the relationships among the observations and variables.

2.6.8 Score plots and loading plots

Lifting out the plane and placing it on the computer screen then produces the plot shown in Figure 2.14. This is the *score plot* and it shows how the observations (countries) are projected from 20 down to two dimensions. We discussed in Chapter 1 how to interpret this picture (see Section 1.5). The other plot shown in Figure 2.15 is known as the *loading plot*, and it displays the relationships among the variables. The loadings are the weights combining the original variables to form the scores. Geometrically, they represent the direction of the plane in the K-space. The score and loading plots are complementary and superimposable, which means that an interesting pattern seen in the score plot can be interpreted by looking along that interesting direction in the loading plot.

The loading plot shows the orientation of the obtained plane ("window") in relation to the original variables. Hence, the plot marks of the variables can be thought of as reflecting the direction and extension of each original variable, as perceived when looking at them from the model plane. This is symbolized by the arrows drawn in Figure 2.15. Arrows have only been indicated for a few variables, to avoid cluttering the plot. Variables that are most important for the model are found on the periphery of the loading plot. Conversely, uninfluential variables are encountered around the origin of the plot (0,0).

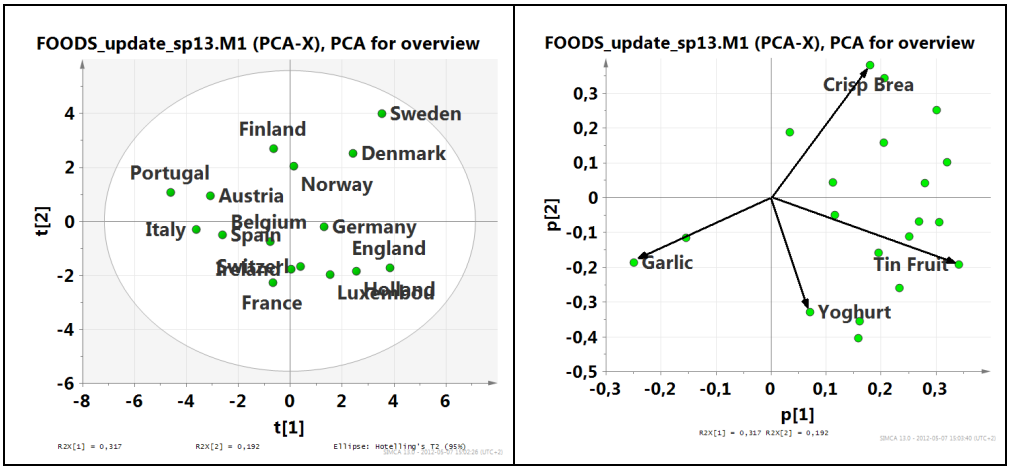


Figure 2.14: (left) PCA score plot of the FOODS example. Figure 2.15: (right) PCA loading plot of the FOODS example. See discussion in Section 1.5.1.

2.6.9 Summary of principles of projections

As seen in the above example, the principles of multivariate projections are simple. On each observation, the values of *K* variables have been measured. These data are represented as points in a multivariate K-dimensional space. One then converts the data table to plots by projecting the data onto a line, a plane, or a hyperplane. The information in the data is seen as patterns along these lines, planes, and hyperplanes, e.g., time trends, clusters, or deviating observations. The direction of the projection plane gives information about which variables are important and which are not, and how the important variables combine to separate the clusters of observations, or to define trends among the observations over time.

Moreover, in the quantitative analysis of relations between variable blocks X and Y we need two spaces, one for the X-variables and one for the Y-variables. Methods like PLS are then employed to develop models that connect the two spaces. But still, these models can be seen as lines and planes in the X- and Y-spaces, with an interpretation very similar to that of PCA.

In summary, with multivariate projection techniques, the measured variables are summarized by the calculation of new variables, called scores. These new variables are embedded in the measured ones, and are therefore referred to as *latent variables*. Using them to create score plots gives us views of the configuration of the observations, "maps".

2.7 The model concept and theoretical basis for model building

The insertion of a window into K-dimensional space is equivalent to fitting a mathematical model to the data. In this section, we will discuss the concept of a model, and the process of model building, mainly with respect to predictive X/Y regression modeling.

2.7.1 The model concept

A model is an approximation of parts of the real world, which simplifies the study of reality. A model is never 100% perfect, but is still very useful. Some simple examples are given in Figure 2.16. Consider the electrified toy train, something which most children have played with. Such a toy train mimics the properties of a full-sized train, though in a miniature format. A map of Iceland, with volcanoes and glaciers marked, is a useful model of reality for tourists. And the score plot of how the European countries relate to each other with respect to patterns in food consumption profiles is yet another model.

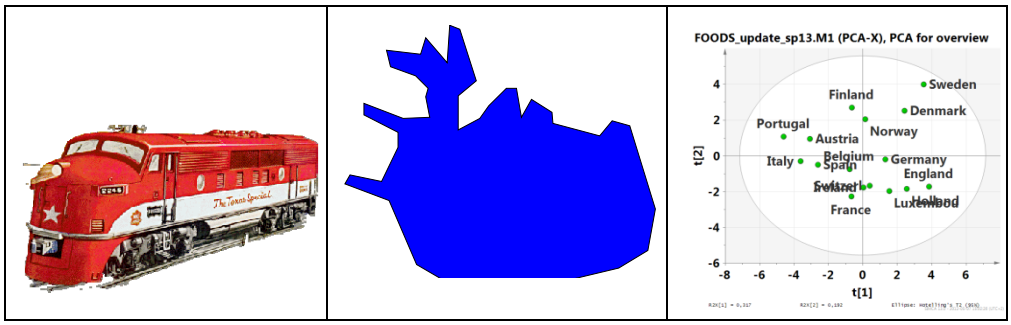


Figure 2.16: Three examples of models, a toy train, a map of Iceland, and a PCA score plot.

Models are not reality, but approximate representations of some aspects of reality. Provided that a model is sound – there are tools to test this – it constitutes an excellent tool for understanding important mechanisms of reality, and for manipulating parts of the reality towards a desired state.

2.7.2 Empirical, semi-empirical, and theoretical models

In this book we work with mathematical models, which can be classified as empirical, semi-empirical, or theoretical models (Figure 2.17). A theoretical model, also called a hard or fundamental model, is usually derived from a well-established and accepted theory. An example is the Schrödinger equation. Theoretical models are often regarded as fundamental "laws" of science - though the label "models" would be more appropriate in many situations. In most cases, however, the mechanism of a system or a process is insufficiently understood, or may be too complicated, to permit a model to be postulated from theory. In such circumstances, an empirical model based on experiments is a useful alternative.

Often some prior knowledge is available to suggest the form the model should take. For instance, the concentration of hydrogen ion in water is usually best represented as the negative logarithm of this concentration, i.e., pH. Or perhaps the experimenters know that the factors A and B are not interesting *per se*, but rather the ratio B/A is what counts. When using this kind of prior knowledge, we construct a semi-empirical model. MVDA and DOE can thus be said to involve semi-empirical modeling of the relationships between factors and responses.

• empirical	• semi-empirical	• fundamental
$y = a + bx + \varepsilon$	$y = a + b\log x + \varepsilon$	$H\psi = E\psi$

Figure 2.17: Examples of empirical, semi-empirical and fundamental mathematical models.

2.7.3 Semi-empirical modeling – Taylor series expansions

Taylor series expansions are often used in semi-empirical modeling. Consider Figure 2.18 in which the "true" relationship between one response and one factor, the function $y = f(x)$, is plotted. We will try to find an alternative function, $y = P(x)$, which can be used to approximate the true function, $y = f(x)$. In a limited interval, Δx, any continuous and differentiable function, $f(x)$, can be arbitrarily well approximated by a polynomial, $P(x)$, giving a Taylor series of the form: $y = P(x) = b_0 + b_1x + b_2x^2 + ... + b_Px^p + e$, where e represents a residual term. In this polynomial function, the degree, p, gives the complexity of the equation.

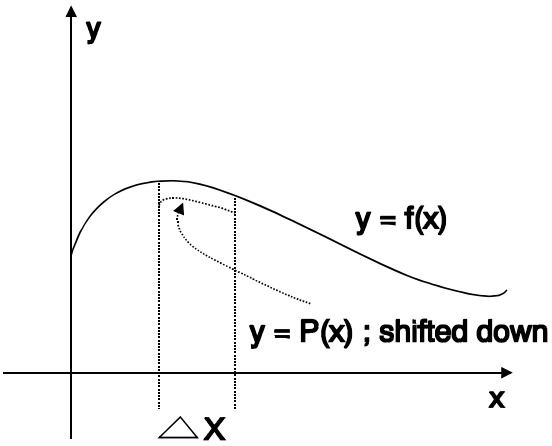


Figure 2.18: The "true" relationship, y = f(x), between a response, y, and a factor x, approximated by an alternative function, y = P(x), in a limited factor interval Δx.

Assume that the degree p is fixed. Then, for this degree, p, the approximation of $y = f(x)$ by $y = P(x)$ gets better as the interval, Δx, becomes smaller. Analogously, for a fixed size of the factor interval, Δx, the approximation is better as the degree, p, becomes higher. Hence, there is a trade-off between model complexity and model range. This kind of reasoning can be generalized to functions of many factors and many responses. We also note that there are infinitely many serial expansions $g(x) = c_0 + c_1g_1(x) + c_2g_2(x) + \ldots$ that approximate f(x) equally well as P(x).

Interestingly, our long experience in technology, chemistry, biology, and medicine, shows that within a restricted experimental area nature is fairly smooth rather than rugged (see Figure 2.19). This means that when several factors are explored, a smooth response surface is typically encountered. Such a smooth response surface can be well approximated by a simple polynomial model, usually of quadratic degree (p = 2).

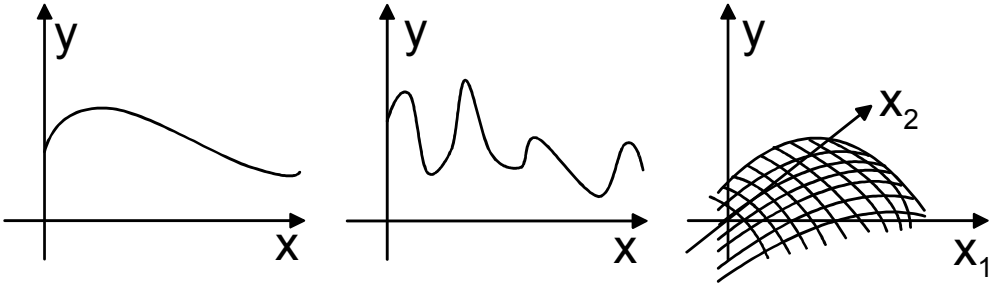


Figure 2.19: Our long experience (technology, chemistry, biology, medicine,...) shows that nature is fairly smooth and not rugged. With several mathematically independent factors a smooth response surface is usually applicable. Such a smooth surface can be well approximated by a quadratic polynomial model.

2.7.4 Conceptual basis of semi-empirical modeling

A semi-empirical model in latent variables is a local model, which describes the situation within the investigated interval. This is a desirable feature. We do not want a global model if it only provides superficial knowledge of reality. A simple example will explain why. Imagine a tourist wanting to find his way to the Eiffel Tower in Paris. Surely, this person would not look at a globe of the planet, because it represents too global a model, with little useful information of precisely where to find the Eiffel Tower. Not even a map of France is local enough, but a map covering the center of Paris would do the job. This map of Paris represents a local model of a small area of reality, and provides detailed information on where to find the great tourist attraction.

Models of complicated systems or processes in science and technology function in the same way. Local models pertaining to narrowly defined regions provide more detail than models covering large regions. Consequently, it is important to carefully specify the range of a model. One way of doing this is to make sure that a representative set of observations (training set) is used for the model development (see Chapters 7 and 18).

2.7.5 Basic conceptual model of projections

For multivariate projection methods (PCA, PCR, PLS, …) the measured variables are modeled as linear combinations of a small set of latent variables. This is illustrated in Figure 2.20. A process is characterized by six signals (measured variables). In the ensuing multivariate analysis, the information in these six variables is used to estimate a few informative variables, the latent variables.

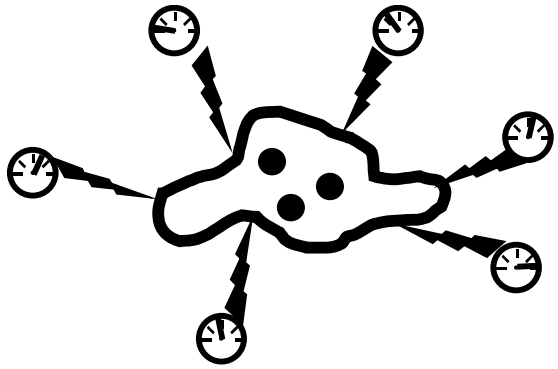


Figure 2.20: Six variables are measured to characterize a process. In the multivariate analysis, these six variables combine to form three estimated latent variables reflecting the state of the process.

The theoretical foundation for the modeling of measured variables by means of latent variables (LV) is based on two principles. The first is that the measurements, by definition, are sums of the underlying latent variables. For example in spectroscopy, the spectrum of a sample is the sum of the spectra of the constituents multiplied by their concentrations in the sample. If the concentrations are t and the spectra p, we get the latent variable model $X = t_1p_1' + t_2p_2' + \ldots = TP' + \text{noise}$. (Beer-Lambert "law"). In many applications this interpretation that the data are explained by a number of "factors" (components) makes sense.

The second theoretical foundation of LV models is one of Taylor expansions (see Section 2.7.3). We assume the data X (and Y in PLS) to be generated by a function F (U,V). Here, the vector variable u describes the change between observations (rows in X) and the vector variable v describes the change between variables (columns in X). If we now make a Taylor expansion of F in the u-direction, after discretizing for i = observation and k = variable, we arrive at the LV model. Again, the smaller the interval of u that is modeled, the fewer terms we need in the Taylor expansion, and the fewer components we need in the LV model. Hence, we can interpret PCA and PLS as models related to similarity. Data (variables) measured on a set of similar observations (samples, items, cases, …) can always be modeled (approximated) by a PCA- or PLS model. And the more similar the observations, the fewer components we need in the model.

We thus have two different interpretations of the LV model. Real data can be interpreted as either a linear combination of "factors" or alternatively as measurements on a set of similar observations. Any mixture of these two interpretations is, of course, also possible.

Latent variables as a concept is more difficult to comprehend, accept, and use, than the straightforward combination of measured variables. In real life, however, we are probably surrounded by many latent variables or "latent concepts", which we discuss and value without even questioning their relevance. One such "latent concept" is the *fitness* of our body. It is hard to measure fitness using only one single numerical value, not only because fitness is a complex phenomenon, but also because fitness is understood and perceived differently by different people. A reasonable description of fitness would probably use information related to body weight, age, heart rate, agility, strength, vigilance, speed, sleeping habits, food consumption, and so on. All these variables would together contribute to the latent variable fitness.

2.7.6 Summary of modeling

Underlying all design and data analysis is the concept of models. This is a consequence of theoretical developments in physics and chemistry by Planck, Bohr, Schrödinger, Heisenberg, and others, and not the least the mathematical - philosophical results of Gödel. Basically, we are forced to realize that complete knowledge about anything is impossible, even in principle, and all we can do is to develop simplified mathematical approximations of the complicated reality, i.e., *models*.

Nevertheless, the use of mathematical models allows us to connect the results of one experiment with those of another, to find favorable directions in the experimental space, in short, to *learn* from experiments. The models used in multivariate data analysis are 'semi-empirical', such as lines, planes, or quadratic surfaces. In cases where good fundamental models exist, they can, of course, be used for the data analysis.

Using the models, the data are separated into two parts; the systematic part "explained" by the model, and the "noise", the remaining unmodeled part of the data. All experiments and measurements have an inherent variability, or "noise".

Data (Y) = Model [factor values (X), parameters (b)] + Noise

From the systematic part of the model, we can now predict the behavior of our system for new values of the X-variables. When the X-variables are factors such as pH or temperature, this allows us to manipulate the system to achieve desired goals and to optimize the system. The noise is modeled by a statistical distribution such as the normal distribution. This allows us to quantify the uncertainty – in terms of confidence intervals. When the X-

variables are signals, and Y measures concentrations or other system properties, we can predict Y from the values of X for new samples by means of a "multivariate standard curve".

2.8 Questions for Chapter 2

1. What are data?
2. What is variability?
3. What is the rationale behind the statement *all data are needed*?
4. Why is multicollinearity a problem for classical regression analysis?
5. Why is it important to distinguish between correlation and causation?
6. What are the three levels of data analysis?
7. What are the advantages of using multivariate projection methods?
8. Are there disadvantages?
9. What is a model?
10. What are the principles of projections?

2.9 Summary and discussion

In the first part of this chapter we discussed the concepts of data and information. It was concluded that data are not synonymous with information, but that data must be processed with analytical tools to highlight the information. When an abundance of data is available, certain challenges arise in the analysis, e.g., dimensionality problems, multicollinearity, noise, and missing data. It was emphasized that all data are needed for the analysis, and that omission of data points without a clear justification should be avoided.

Next, different approaches to the analysis of multivariate data were outlined. The traditional COST approach gives an uninformative and even misleading evaluation, and the use of classical statistics is severely hampered by unrealistic assumptions about the data and the models. Multivariate projection methods are more realistic.

The chapter ended with the model concept and the principles of projections. A model is an approximation of reality that will never be completely perfect, but a good model constitutes an efficient tool for understanding, and sometimes also manipulating, the reality in a desired direction.

With regard to projections, the most fundamental aspect is that a row in a data table can be interpreted as a point in a multidimensional space. This allows multivariate data to be understood and analyzed in terms of spaces and geometrical configurations within these spaces. A projection method can be thought of as inserting a window into a multidimensional space, and then looking at the distribution of the data points when projected onto this window.

3 PCA

3.1 Objective

The objective of the third chapter is to introduce principal component analysis, PCA. This is a multivariate projection method designed to extract and display the systematic variation in a data matrix X. A geometric perspective will be used to give an intuitive understanding of the method. With PCA a number of diagnostic and interpretational tools become available, which will be outlined.

3.2 Introduction to PCA

Principal component analysis forms the basis for multivariate data analysis [Jackson, 1991; Wold, *et al.*, 1984; Wold, *et al*., 1987]. As shown by Figure 3.1, the starting point for PCA is a matrix of data with *N* rows (*observations*) and *K* columns (*variables*), here denoted by X. The observations can be analytical samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals, trials of a DOE-protocol, and so on. In order to characterize the properties of the observations, one measures variables. These variables may be of spectral origin (NIR, NMR, IR, UV, X-ray, …), chromatographic origin (HPLC, GC, TLC, …), or they may be measurements from sensors in a process (temperatures, flows, pressures, curves, etc.).

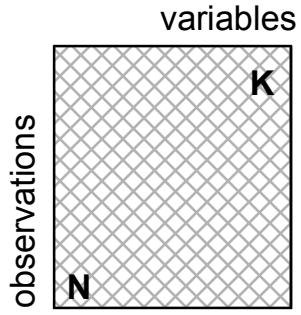


Figure 3.1: Notation used in PCA. The observations (rows) can be analytical samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals, trials of a DOE-protocol, and so on. The variables (columns) might be of spectral origin, of chromatographic origin, or be measurements from sensors and instruments in a process.

PCA goes back to Cauchy, but was first formulated in statistics by Pearson, who described the analysis as finding *lines and planes of closest fit to systems of points in space* [Jackson, 1991]. The most important use of PCA is indeed to represent a multivariate data table as a low-dimensional plane, usually consisting of 2 to 5 dimensions, such that an overview of the data is obtained. This overview may reveal groups of observations, trends, and outliers. This overview also uncovers the relationships between observations and variables, and among the variables themselves.

Statistically, PCA finds lines, planes and hyperplanes in the K-dimensional space that approximate the data as well as possible in the least squares sense. It is easy to see that a line or a plane that is the least squares approximation of a set of data points makes the variance of the co-ordinates on the line or plane as large as possible (Figure 3.2).

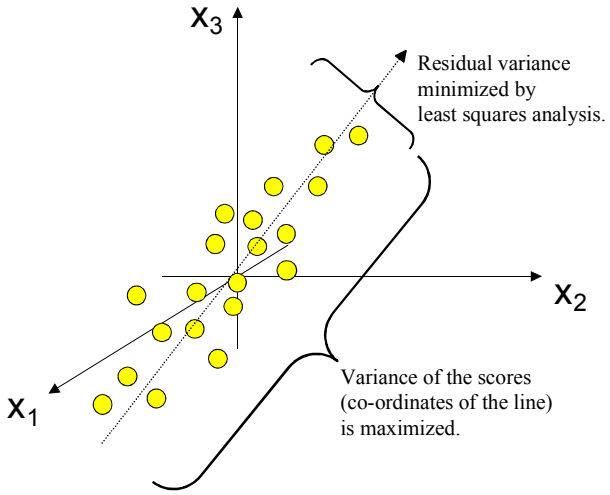


Figure 3.2: PCA derives a model that fits the data as well as possible in the least squares sense. Alternatively, PCA may be understood as maximizing the variance of the projection co-ordinates.

3.3 Pre-treatment of data

Prior to PCA, data are often pre-treated, in order to transform the data into a form suitable for analysis, i.e., to re-shape the data such that important assumptions are better fulfilled. In fact, pre-processing can make the difference between a useful model and no model at all. In this section *scaling of data* and *mean-centering* are described. Additional pre-processing tools like *transformations* (Chapter 12), *advanced scaling* (Chapter 13), and *data correction and compression* (Chapter 14) are addressed later.

3.3.1 Scaling

Variables often have substantially different numerical ranges. A variable with a large range has a large variance, whereas a variable with a small range has a small variance. Since PCA is a maximum variance projection method, it follows that a variable with a large variance is more likely to be expressed in the modeling than a low-variance variable. As an example, consider the LOWARP dataset (Table 3.1), and particularly the wrp3 and st4 variables. Here, wrp3 varies between 0.2 and 1.0, whereas st4 ranges from roughly 17000 to 30000. As a consequence, st4 will dominate over wrp3, unless the data are normalized.

Table 3.1: The LOWARP dataset.

		1	2	3	4	5	6	7	8	9	10	13	16	11	12	14	15	17	18
Num	Name	glas	crtp	mica	amtp	wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	wrp7	wrp8	st1	st2	st3	st4	st5	st6
1	1	40	10	10	40	0.9	5	0.2	1	0.3	4.2	1.2	1.3	232	15120	2190	26390	2400	0.7
2	2	20	20	0	60	3.7	7.3	0.7	1.8	2.5	5.4	1.8	2.1	150	12230	905	20270	1020	0.6
3	3	40	20	0	40	3.6	6.9	0.9	2.1	4.8	9.4	1.2	1.4	243	15550	1740	21180	1640	
4	4	20	20	20	40	0.6	3.1	0.3	0.4	0.4	1.1	1	1	188	11080	1700	17630	1860	0.5
5	5	20	10	20	50	0.3	2.1	0.3	0.3	0.8	1.1	1.2	1.3	172	11960	1810	21070	1970	0.5
6	6	40	0	20	40	1.2	5					1.1	1.3	245	15600	2590	25310	2490	0.6
7	7	20	0	20	60	2.3	3.9	0.3	0.4	0.7	1.4	1.5	1.6	242	13900	1890	21370	1780	
8	8	40	0	10	50	2.6	5.9	0.4	0.2	0.7	1.2	1.6	1.6	243	17290	2130	30530	2320	0.7
9	9	30	20	10	40	2.2	5.3	0.2	0.7	0.6	2	1	1.1	204	11170	1670	19070	1890	0.6
10	10	40	0	0	60	5.8	7	0.9	1	5.6	11.8	1.6	1.8	262	20160	1930	29830	1890	
11	11	30	0	20	50	0.8	2.9	0.5	0.6	1.1	2	1.3	1.3	225	14140	2140	22850	2110	0.7
12	12	30	10	0	60	2.8	5.1	1	1.2	2.7	6.1	1.9	2.1	184	15170	1230	23400	1250	0.6
13	13	30	10	10	50	1.1	4.7	0.6	0.9	1.3	3.5	1.4	1.4	198	13420	1750	23790	1930	0.7
14	14	30	10	10	50	1.9	4.7	1	1	2.8	5.4	1.5	1.6	234	16970	1920	25010	1790	0.7
15	15	30	10	10	50	2.9	5.9	0.5	0.6	1	6.6	1.5	1.6	239	15480	1800	23140	1730	
16	16	40	10	0	50	5.5	7.9	0.8	2.4	5.5	9.3	1.5	1.8	256	18870	1880	28440	1790	
17	17	30	0	10	60	3.2	6	0.3	0.5	1.5	5.2	1.5	1.7	249	16310	1860	24710	1780	

Another simple example will further illustrate the concept of, and need for, scaling. In connection with a pre-season friendly game of football (soccer), the trainers of both teams decided to measure the body weight (in kg) of their players. The trainers also recorded the body height (in m) of each player. These data are listed in Table 3.2 and plotted in two ways in Figures 3.3 and 3.4.

Table 3.2: Measured body weights and body heights of 23 individuals.

Height (m)	**1.8**	**1.61**	**1.68**	**1.75**	**1.74**	**1.67**	**1.72**	**1.98**	**1.92**	**1.7**	**1.77**	**1.92**
Weight (kg)	**86**	**74**	**73**	**84**	**79**	**78**	**80**	**96**	**90**	**80**	**86**	**93**
Height (m)	**1.6**	**1.85**	**1.87**	**1.94**	**1.89**	**1.89**	**1.86**	**1.78**	**1.75**	**1.8**	**1.68**	
Weight (kg)	**75**	**84**	**85**	**96**	**94**	**86**	**88**	**99**	**80**	**82**	**76**	

When the two variables are plotted in a scatter plot where each axis has the same scale – the x and y axes both extend over 30 units – we can see that the data points *only* spread in the vertical direction (Figure 3.3). This is because body weight has a much larger numerical range than body height. Should we analyze these data with PCA, without any pre-processing, the results would only reflect the variation in body weight.

Actually, this dataset contains an atypical observation (individual). This is much easier to see when the two variables are more appropriately scaled (Figure 3.4). Here, we have compressed the variation along the body weight axis and zoomed in on body height. There is a strong correlation between body height and body weight, except for one outlier in the data. This was impossible to see in the previous plot when body weight dominated over body height. We have therefore scaled the data such that both variables make the same contribution to the model.

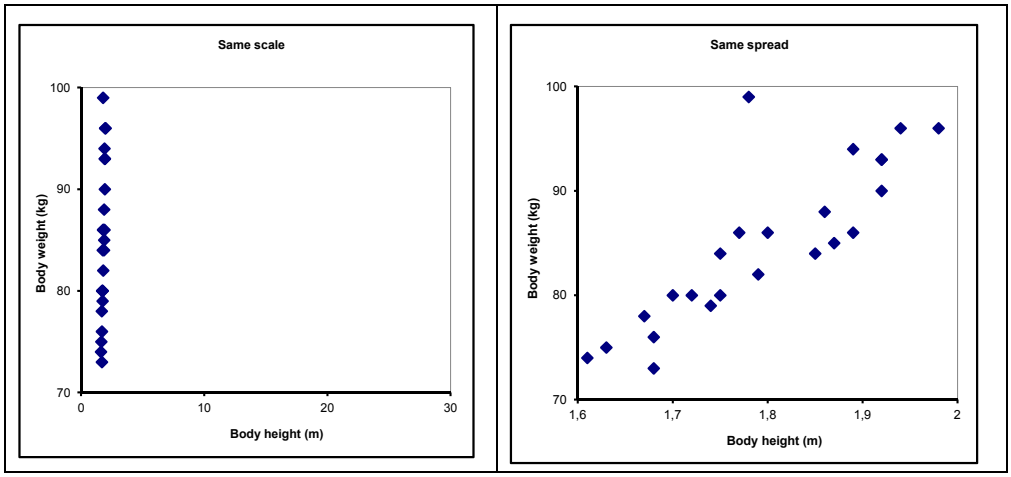


Figure 3.3: (left) Scatter plot of body weight versus body height of 23 individuals. The data pattern is dominated by the influence of body weight. The variables have the same scale. Figure 3.4: (right) Scatter plot of body weight against body height of 23 individuals. Now, the variables are given equal importance by displaying them according to the same spread. An outlier, a deviating individual, is now discernible. We have good reasons to believe (admittedly, based on some detailed insight of one of the Umetrics' staff members) that the outlying observation (the atypical person) happens to be the referee of the game.

In order to give both variables, body weight and body height, equal weight in the data analysis, we standardized them. Such a standardization is also known as "scaling" or "weighting", and means that the length of each co-ordinate axis in the variable space is regulated according to a pre-determined criterion (Figure 3.5). The first time a dataset is analyzed it is recommended to set the length of each variable axis to equal length.

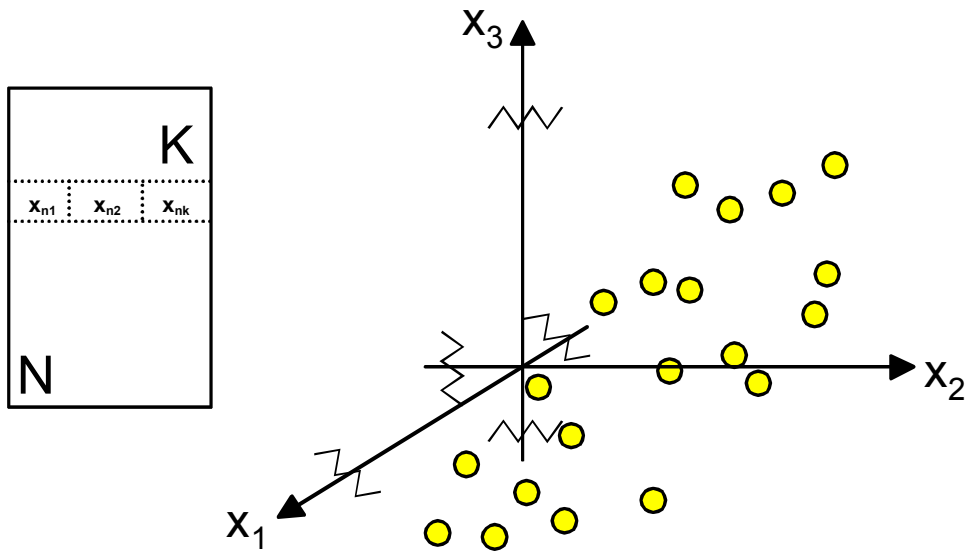


Figure 3.5: The scaling of variables means that the length of each co-ordinate axis in the variable space is regulated according to some selected criterion. The most common criterion is that the length of each variable axis be set to be the same variance.

There are many ways to scale the data, but the most common technique is the *unit variance (UV) scaling*. For each variable (column) one calculates the standard deviation (s_k) and obtains the scaling weight as the inverse standard deviation ($1/s_k$). Subsequently, each column of **X** is multiplied by $1/s_k$. Each scaled variable then has equal (unit) variance. Another name for this scaling method is "auto-scaling".

A simple geometrical understanding of UV-scaling is based on the equivalence between the length of a vector and its standard deviation (square root of variance). Hence, the initial variance of a variable is interpretable as the squared "size" or "length" of that variable. This means that with UV-scaling we accomplish a shrinking of "long" variables and a stretching of "short" ones (Figure 3.6). By putting all variables on a comparable footing, no variable is allowed to dominate over another because of its length.

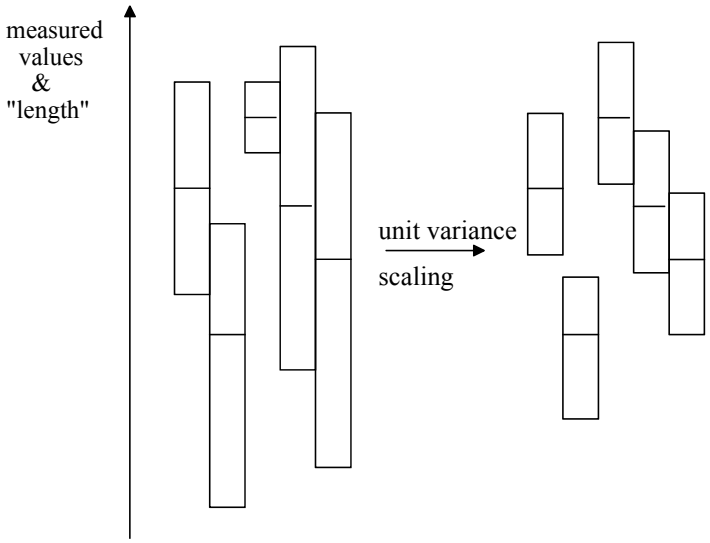


Figure 3.6: The effect of unit variance scaling. The vertical axis represents the "length" of the variables and their numerical values. Each bar corresponds to one variable and the short horizontal line inside each bar represents the mean value. Prior to any pre-processing the variables have different variances and mean values. After scaling to unit variance, the "length" of each variable is identical. The mean values still remain different, however.

Like any projection method, PCA is sensitive to scaling. This means that by modifying the variance of the variables, it is possible to attribute different importance to them. This gives the possibility of down-weighting irrelevant or noisy variables. However, one must not overlook the risk of scaling subjectively to give you the model you want. Generally, UV-scaling is the most objective approach, and is recommended if there is no prior information about the data. Sometimes no scaling at all would be appropriate, especially with data where all the variables are expressed in the same unit, for instance, with spectroscopic data. Later on, when more experience has been gained, more elaborate scaling procedures may be used.

3.3.2 Mean-centering

Mean-centering is the second part of the standard procedure for pre-preprocessing. With mean-centering the average value of each variable is calculated and then subtracted from the data. This improves the interpretability of the model. A graphical interpretation of mean-centering is shown in Figure 3.7.

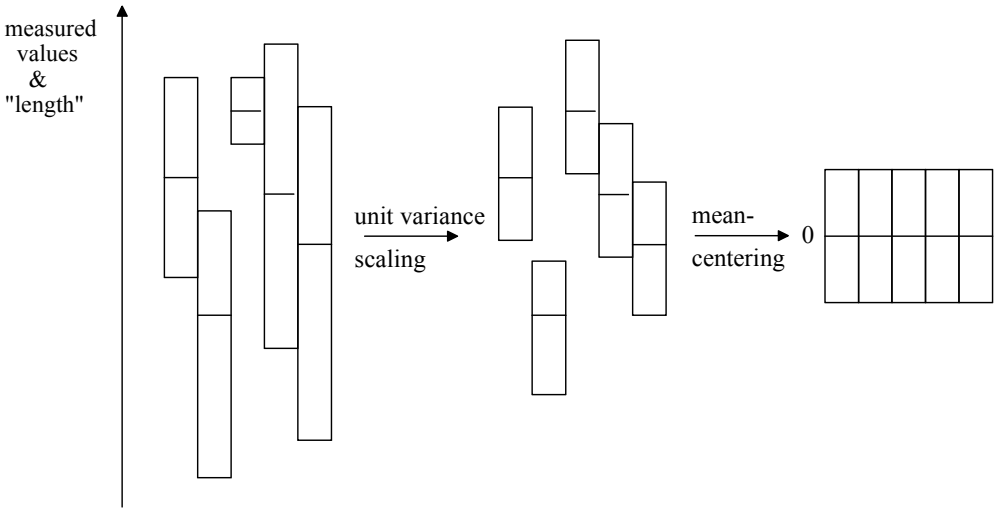


Figure 3.7: After mean-centering and unit variance scaling all variables will have equal "length" and mean value zero.

The mean-centering and UV-scaling procedures are applied by default in SIMCA. Note, however, that in some cases, such as multivariate calibration, it is not necessarily advantageous to use this combination of pre-processing tools, and some other choice might be more appropriate.

3.4 A geometric interpretation of PCA

We will now explain how PCA works: initially, using a geometrical approach, followed by a more formal algebraic account.

3.4.1 Setting up K-dimensional space

Consider a matrix **X** with *N* observations and *K* variables. For this matrix we construct a variable space with as many dimensions as there are variables (Figure 3.8). Each variable represents one co-ordinate axis. For each variable, the length has been standardized according to a scaling criterion, normally by scaling to unit variance.

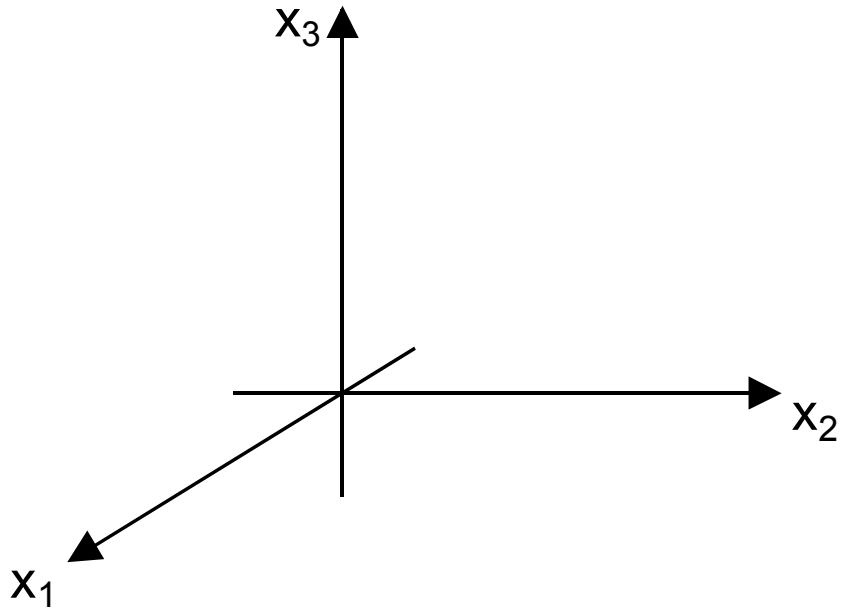


Figure 3.8: A K-dimensional variable space. For simplicity, only three variable axes are displayed. The "length" of each co-ordinate axis has been standardized according to a specific criterion, usually unit variance scaling.

3.4.2 Plotting the observations in K-dimensional space

In the next step, each observation (each row) of the X-matrix is placed in the K-dimensional variable space. Consequently, the rows in the data table form a swarm of points in this space (Figure 3.9).

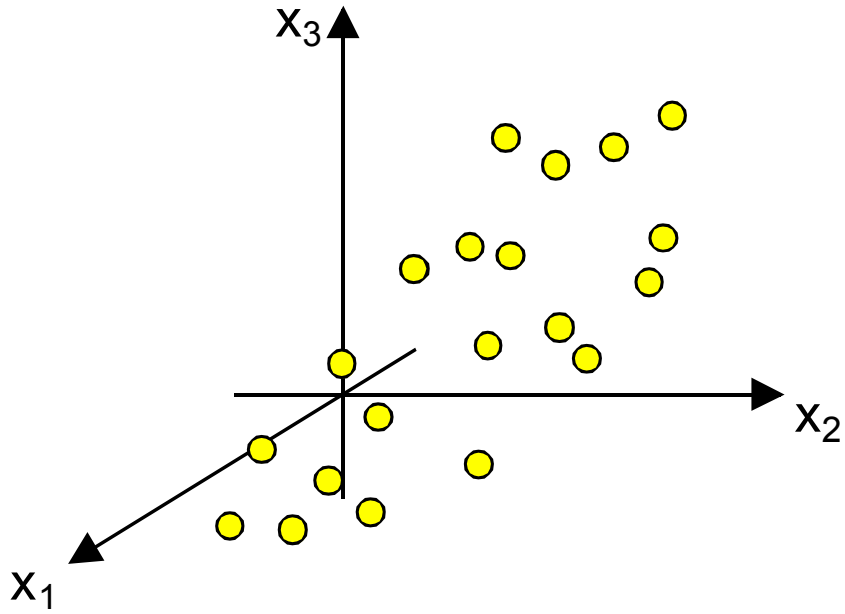


Figure 3.9: The observations (rows) in the data matrix X can be understood as a swarm of points in the variable space (K-space).

3.4.3 The effect of mean-centering

The mean-centering involves the subtraction of the variable averages from the data. This vector of averages corresponds to a point in the K-space (Figure 3.10).

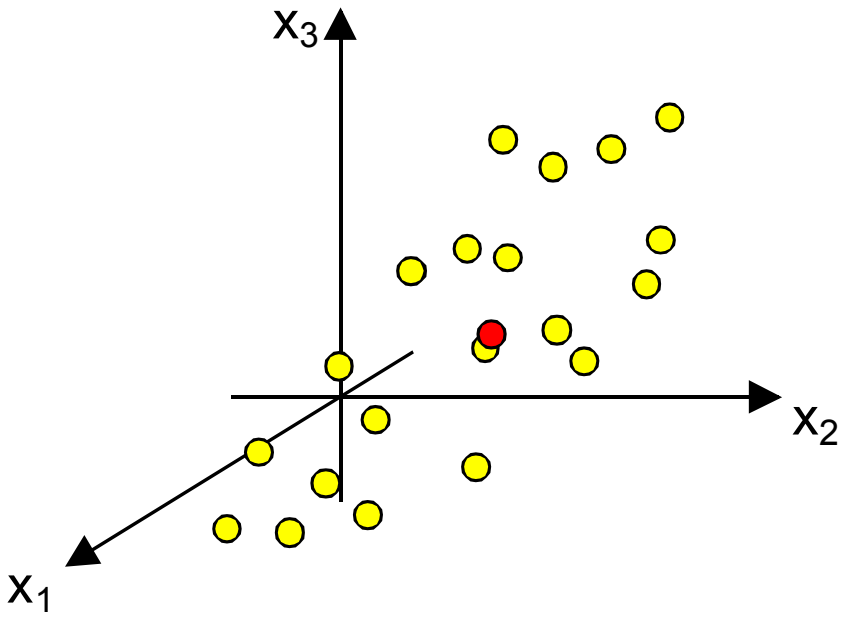


Figure 3.10: In the mean-centering procedure one first computes the variable averages. This vector of averages is interpretable as a point (here: in red) in space. This point is situated in the middle of the point swarm (at the center of gravity).

The subtraction of the averages from the data corresponds to a re-positioning of the co-ordinate system, such that the average point now is the origin (Figure 3.11).

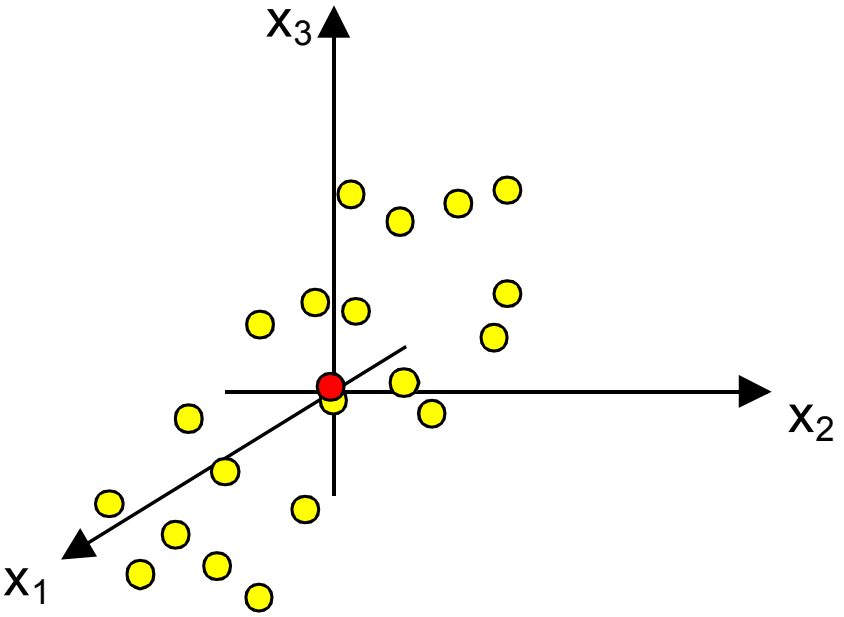


Figure 3.11: The mean-centering procedure corresponds to moving the origin of the co-ordinate system to coincide with the average point (here: in red).

3.4.4 The first principal component

After mean-centering and scaling to unit variance the dataset is ready for the computation of the first principal component (PC1). This component is the line in the K-dimensional space that best approximates the data in the least squares sense. This line goes through the average point (Figure 3.12). Each observation may now be projected onto this line in order to get a co-ordinate value along the PC-line. This new co-ordinate value is known as a *score*.

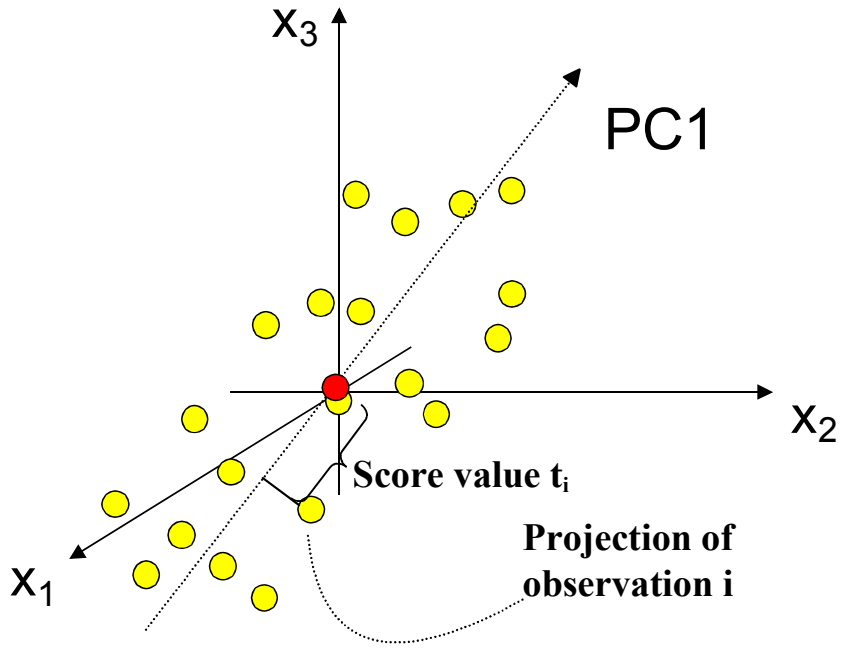


Figure 3.12: The first principal component, PC1, is the line which best accounts for the shape of the point swarm. It represents the maximum variance direction in the data. Each observation may be projected onto this line in order to get a co-ordinate value along the PC-line. This value is known as a score.

3.4.5 Extending the model with the second principal component

Usually, one principal component is insufficient to model the systematic variation of a dataset. Thus, a second principal component, PC2, is calculated. The second PC is also represented by a line in the K-dimensional variable space, which is orthogonal to the first PC (Figure 3.13). This line also passes through the average point, and improves the approximation of the X-data as much as possible.

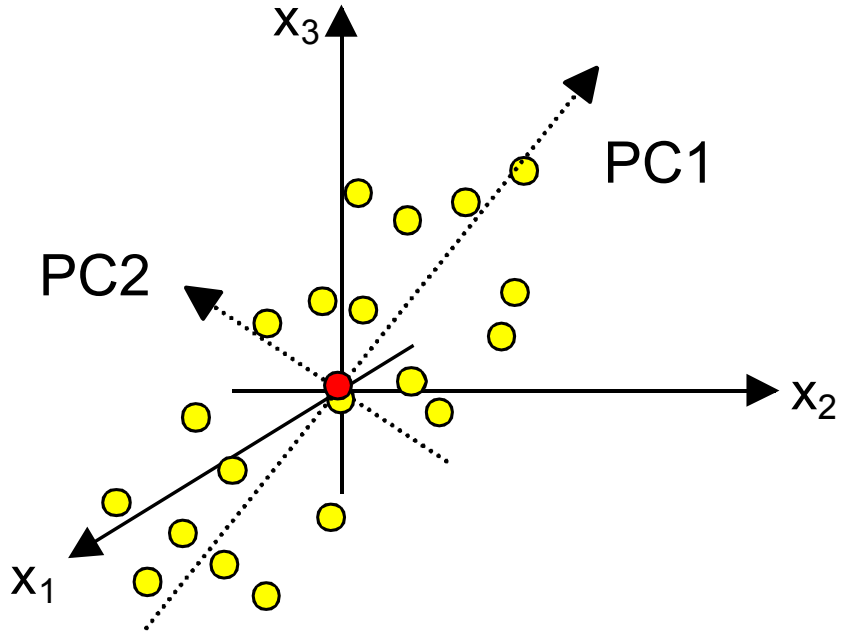


Figure 3.13: The second principal component, PC2, is oriented such that it reflects the second largest source of variation in the data, while being orthogonal to the first PC. PC2 also passes through the average point.

3.4.6 Two principal components define a model plane

When two principal components have been derived, they together define a plane, a window into the K-dimensional variable space (Figure 3.14). By projecting all the observations onto this low-dimensional sub-space and plotting the results, it is possible to visualize the structure of the investigated dataset. The co-ordinate values of the observations on this plane are called *scores*, and hence the plotting of such a projected configuration is known as a *score plot*.

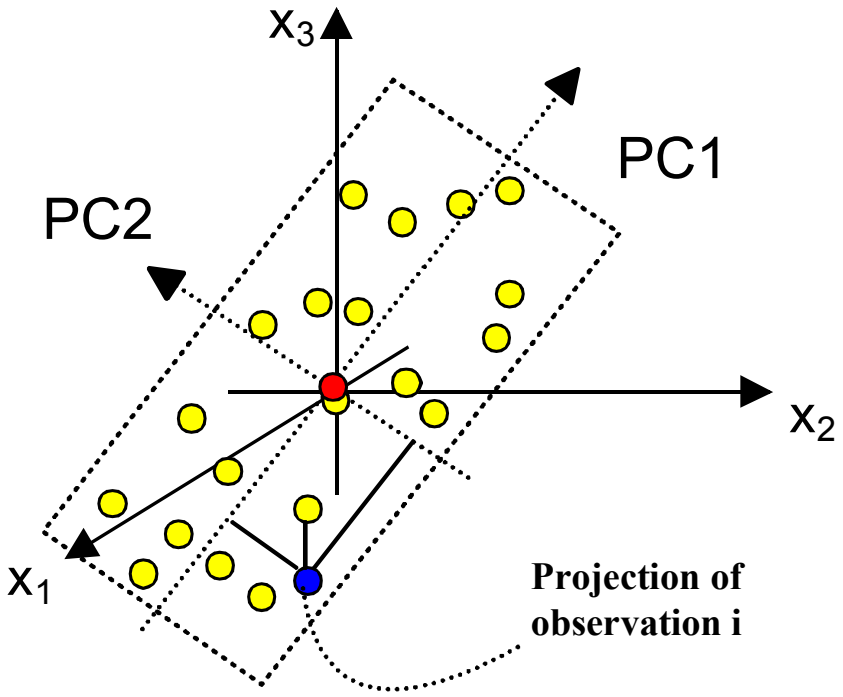


Figure 3.14: Two PCs form a plane. This plane is a window into the multidimensional space, which can be visualized graphically. Each observation may be projected onto this, giving a score for each.

Let us now reconsider the FOODS dataset. Figure 3.15 displays the plot obtained when plotting the scores of the two first principal components. These scores are called t_1 and t_2 (this notation is better explained in Section 3.4.9).

As seen in Figure 3.15, each European country is characterized by two values, one along the first PC and another along the second PC. This score plot is a map of the 16 countries. Countries close to each other have similar properties, whereas those far from each other are dissimilar with respect to food consumption profiles. The Nordic countries (Finland, Norway, Denmark and Sweden) are located together in the upper right-hand corner, thus representing a group of nations with some similarity in food consumption. Belgium and Germany are close to the center (origin) of the plane, which indicates that they have average properties.

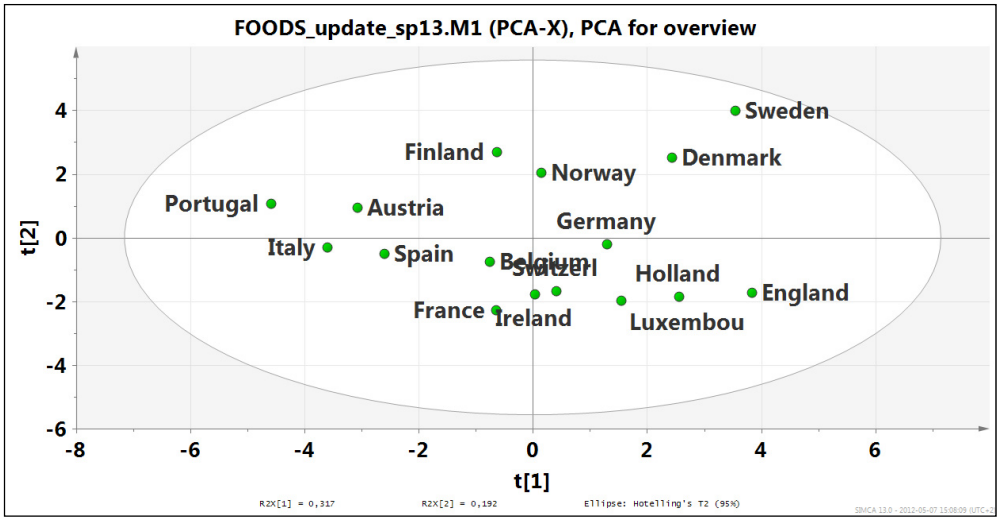


Figure 3.15: PCA score plot of the two first PCs of the FOODS dataset. This provides a map of how the countries relate to each other. The first component explains 32% of the variation and the second component 19%.

3.4.7 How to interpret the score plot

In a PCA model with two components, that is, a plane in K-space, we wonder which variables are responsible for the patterns seen among the observations? We would like to know which variables are influential, and also how the variables are correlated. Such knowledge is given by the principal component *loadings* (Figure 3.16). These loading vectors are called p_1 and p_2 (see further discussion in Section 3.4.9).

Figure 3.16 displays the relationships between *all* 20 variables at the same time. Variables contributing similar information are grouped together: that is, they are correlated. Crisp bread (Crisp_Br) and frozen fish (Fro_Fish) are examples of two variables which are positively correlated. When the numerical value of one variable increases or decreases, the numerical value of the other variable has a tendency to change in the same way.

When variables are negatively ("inversely") correlated they are positioned on opposite sides of the plot origin, in diagonally opposed quadrants. For instance, the variables garlic and sweetener are inversely correlated, meaning that when garlic increases, sweetener decreases, and vice versa.

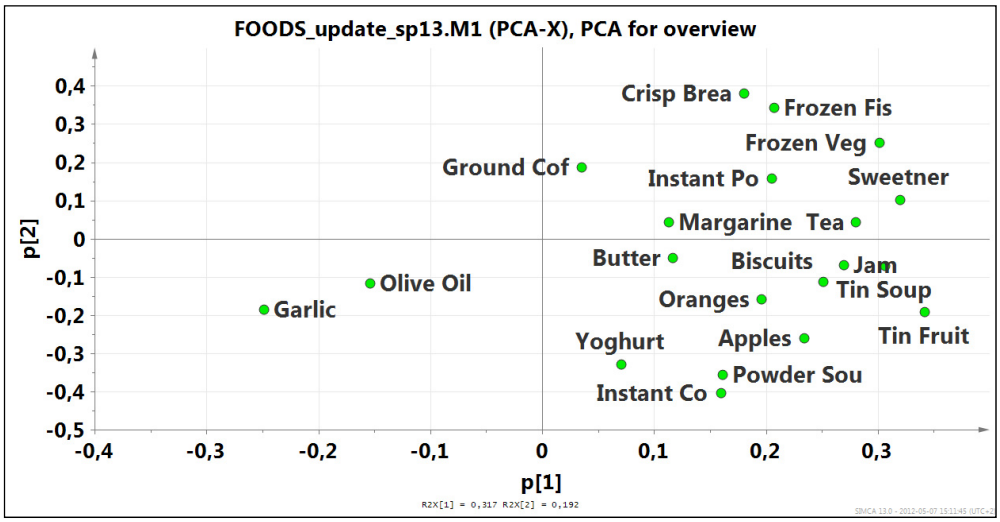


Figure 3.16: PCA loading plot of the first two principal components (p_2 vs. p_1) of the FOODS dataset.

Furthermore, the *distance* to the origin also conveys information. The further away from the plot origin a variable lies, the stronger impact that variable has on the model. This means, for instance, that the variables crisp bread (Crisp_Br), frozen fish (Fro_Fish), frozen vegetables (Fro_Veg) and garlic (Garlic) separate the four Nordic countries from the others. The four Nordic countries are characterized by having high values (high consumption) of the former three provisions and low consumption of garlic. Moreover, the model interpretation suggests that countries like Italy, Portugal, Spain, and to some extent Austria, have high consumption of Garlic, and low consumption of sweetener (Sweetener), tinned soup (Ti_Soup) and tinned fruit (Ti_Fruit).

Geometrically, the principal component loadings express the orientation of the model plane in the K-dimensional variable space (Figure 3.17). The direction of PC1 in relation to the original variables is given by the cosine of the angles α_1, α_2, and α_3. These values indicate how the original variables x_1, x_2, and x_3 "load" into (= contribute to) PC1. Hence, they are called loadings. Of course, a second set of loading coefficients expresses the direction of PC2 in relation to the original variables. Hence, with two PCs and three original variables, six loading values (cosine of angles) are needed to specify how the model plane is positioned in the K-space.

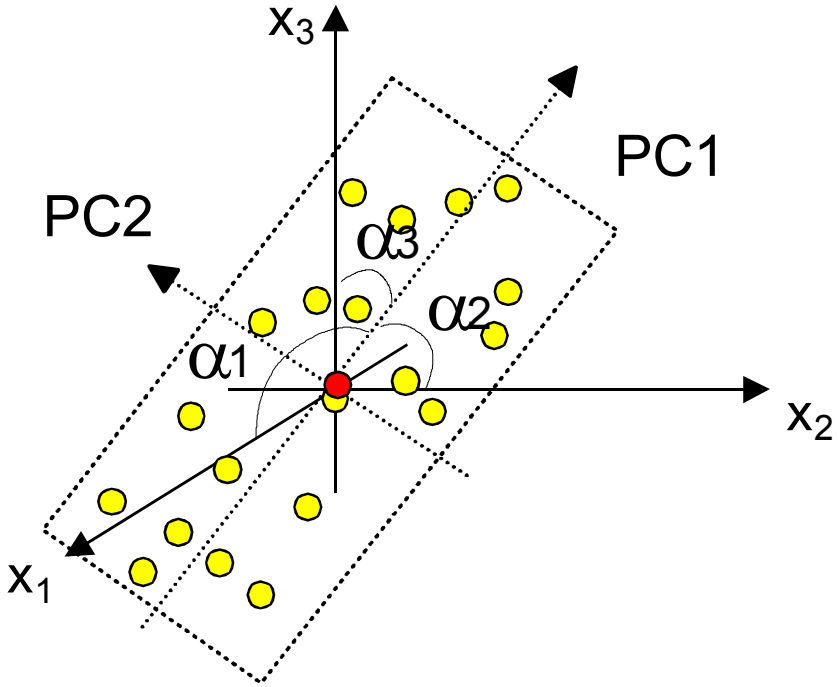


Figure 3.17: The principal component loadings uncover how the PCA model plane is inserted in the variable space. The loadings are used for interpreting the meaning of the scores.

3.4.8 Extensions to higher-order components

Frequently, one or two principal components are not enough to adequately summarize the information in a dataset. In such cases, the descriptive ability of the PCA model improves by using more principal components. There are several approaches that can be used to evaluate how many principal components are appropriate [Jackson, 1991].

Building on the two-dimensional PCA model of the FOODS dataset, we computed a third principal component. This third PC (i) is oriented in the direction of the third largest variation in the data, (ii) is orthogonal to the other two, and (iii) passes through the average point (the origin). The orthogonality constraint thus means that the third component becomes perpendicular to the already existing model plane.

The scores of the third and first PCs (t_3 vs. t_1) are plotted in Figure 3.18. Thus, in comparison with Figure 3.15, the vertical axis has been changed, while the horizontal axis is preserved. The most striking feature in this score plot (Figure 3.18) is the deviating behavior of England and Ireland.

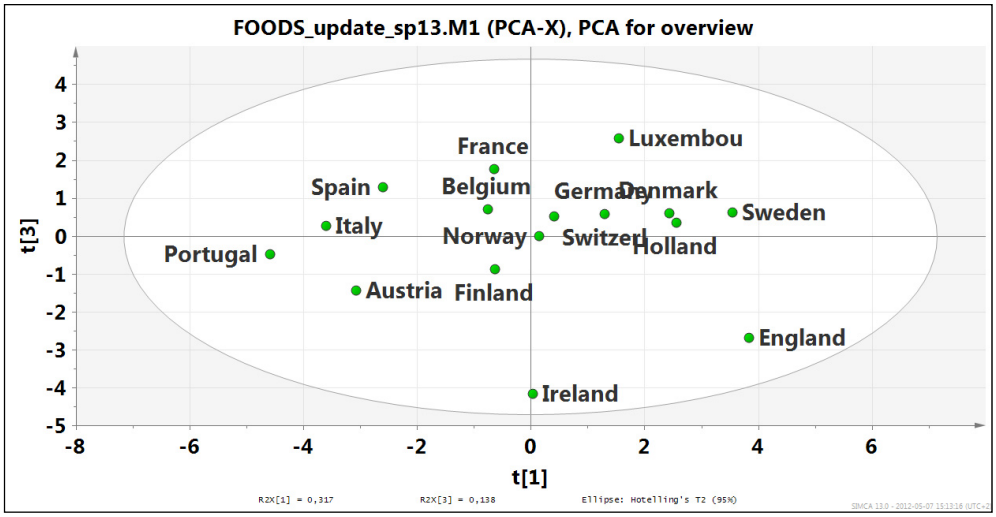


Figure 3.18: PCA score plot of the third PC (t_3) versus the first PC (t_1) of the FOODS dataset. The first component explains 32% of the variation and the third component 14%.

The scores of the third PC are accompanied by the corresponding loadings. A scatter plot of the loadings of the third component versus the loadings of the first component (p_3 vs. p_1) is shown in Figure 3.19. This plot indicates that it is mainly the variables tea and jam which govern the positioning of England and Ireland. England especially, but to some extent also Ireland, consumes larger amounts of tea and jam compared with the other countries included in this survey. Additionally, England and Ireland exhibit less than average consumption of garlic and olive oil.

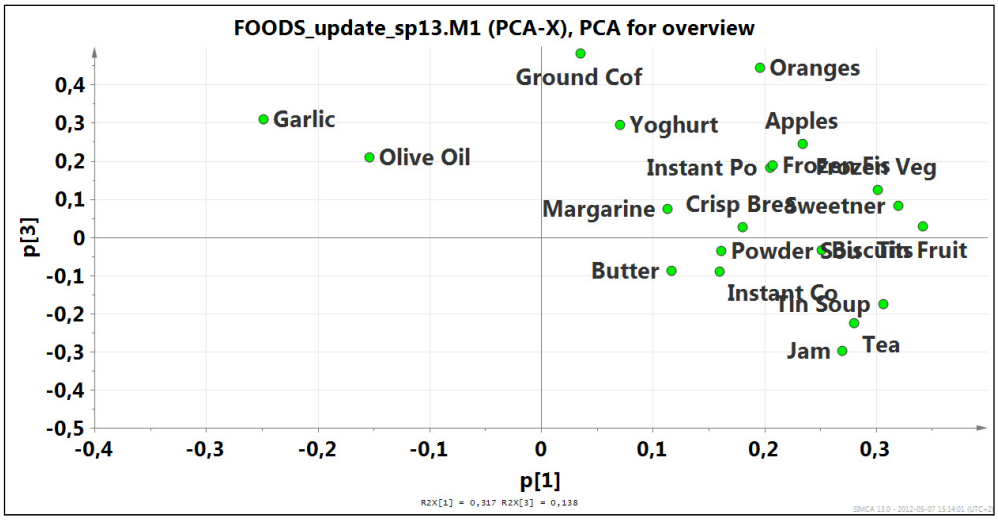


Figure 3.19: Scatter plot of the loadings of the third and first principal components (p_3 vs. p_1) of the FOODS dataset.

3.4.9 Summary of PCA

By using PCA, a data table **X** is modeled as

$$\mathbf{X} = \mathbf{1}*\bar{\mathbf{x}}' + \mathbf{T}*\mathbf{P}' + \mathbf{E} \qquad \text{(eqn. 3.1)}$$

In the expression above, the first term, $\mathbf{1}*\bar{\mathbf{x}}'$, represents the variable averages and originates from the pre-processing step. The second term, the matrix product **T*P´**, models the structure, and the third term, the residual matrix **E**, contains the noise.

The principal component scores of the first, second, third, …, components ($\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, …) are columns of the score matrix **T**. These scores are the co-ordinates of the observations in the model (hyper-)plane. Alternatively, these scores may be seen as new variables which summarize the old ones (Figure 3.20). In their derivation, the scores are sorted in descending importance (t_1 explains more variation than t_2, t_2 explains more variation than t_3, and so on). Typically, 2 to 5 principal components are sufficient to approximate a data table well.

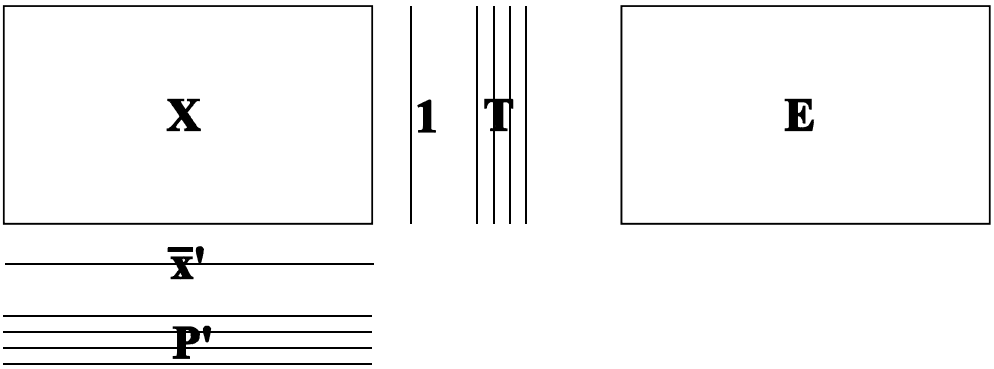


Figure 3.20: A matrix representation of how a data table X is modeled by PCA.

The meaning of the scores is given by the loadings. The loadings of the first, second, third, …, components ($\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$,..) build up the loading matrix **P** (Figure 3.20). Note that in Figure 3.20 a prime has been used with **P** to denote its transpose.

The loadings define the orientation of the PC plane with respect to the original X-variables. Algebraically, the loadings inform how the variables are linearly combined to form the scores. The loadings unravel the *magnitude* (large or small correlation) and the *manner* (positive or negative correlation) in which the measured variables contribute to the scores.

To clarify what is plotted and when, Figure 3.21 provides a final overview of the various scores and loadings plots pertaining to the FOODS example.

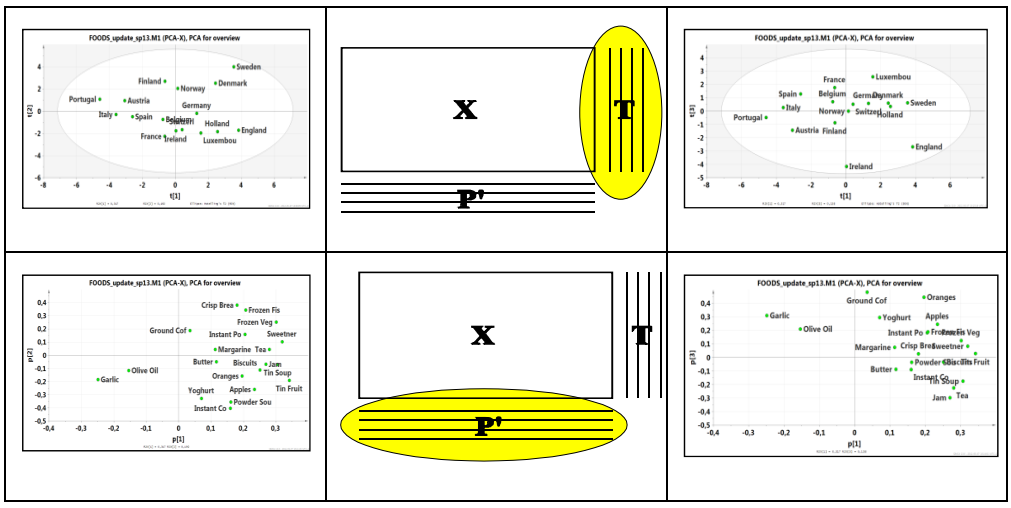


Figure 3.21: An overview of which PCA parameters are plotted in connection with the FOODS dataset.

3.5 Additional PCA diagnostics

PCA offers a number of useful model parameters and other diagnostic tools, which can be displayed graphically or numerically [Wold, *et al.*, 1984; Wold, *et al.*, 1987]. In this section, we will continue the discussion of scores and loadings, but also discuss the *residuals*, the deviations of the real data from the model, and explore some diagnostics related to these. In addition, the technique of cross-validation will be outlined; this assesses the complexity and predictive power of the model.

3.5.1 Observation diagnostics – Are there outliers in the data?

PCA discovers *strong* outliers and *moderate* outliers. Conceptually, outliers are observations that are extreme or that do not fit the PCA model. Outliers are both serious and interesting, but easy to detect. Strong outliers are found in plots of PCA scores and moderate outliers are found by inspecting the model residuals [Wikström, *et al.*, 1998a]. By the term *residuals* we mean the X-variation that was not captured by the PCA model and constitutes the matrix **E** in equation 3.1.

Strong outliers are found in the score plots. They have high *leverage* on the model, i.e., strong "power" to pull the PCA model toward themselves, and may "consume" one PC just because of their existence (Figures 3.22 and 3.23). The term *leverage* derives from the Archimedean principle that anything can be lifted out of balance as long as the lifter has a long enough lever. Leverage is a measure of the influence of an observation and is proportional to the distance of the observation from the center of the data (see details in Appendix II).

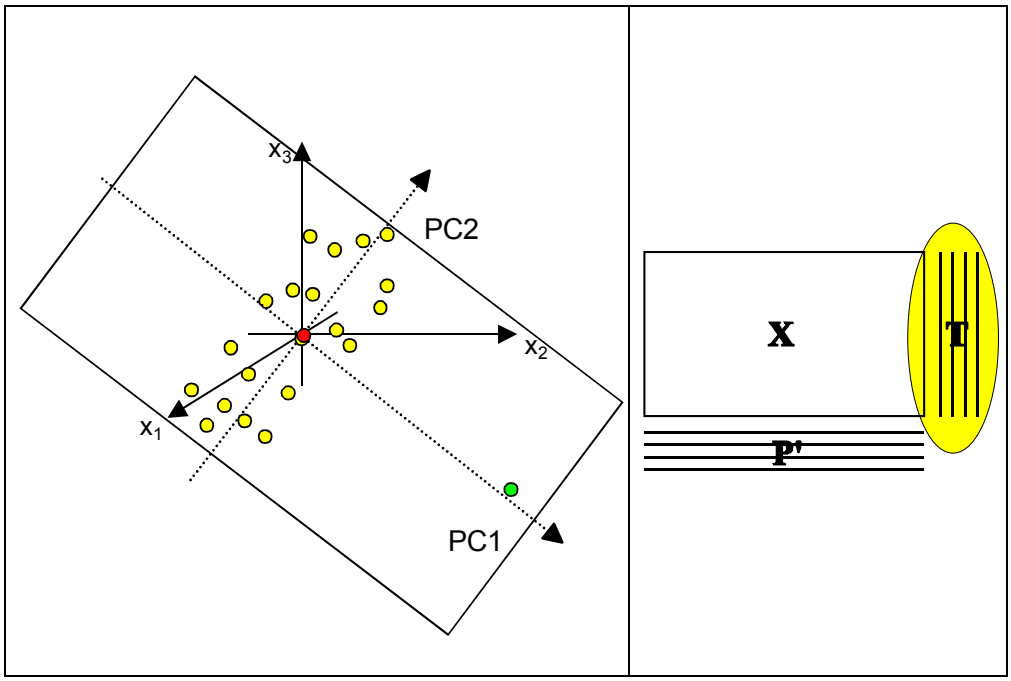


Figure 3.22: (left) Strong outliers are found in score plots. They have high leverage *on the model, i.e., strong "power" to rotate the PCA model towards themselves. Figure 3.23: (right) Plotting of PCA scores is useful for identifying strong outliers.*

A diagnostic showing strong outliers is given by *Hotelling's* T^2 [Jackson, 1991; Wikström, *et al.*, 1998a]. This statistic is a multivariate generalization of the Student's t-test, and provides a check for observations adhering to multivariate normality. A definition of Hotelling's T^2 is given in Appendix II.

When used in conjunction with a score plot, Hotelling's T^2 defines the normal (operating) area corresponding to, for instance, 95% or 99% confidence. Figures 3.24 and 3.25 demonstrate cases where strong outliers are present and absent, respectively.

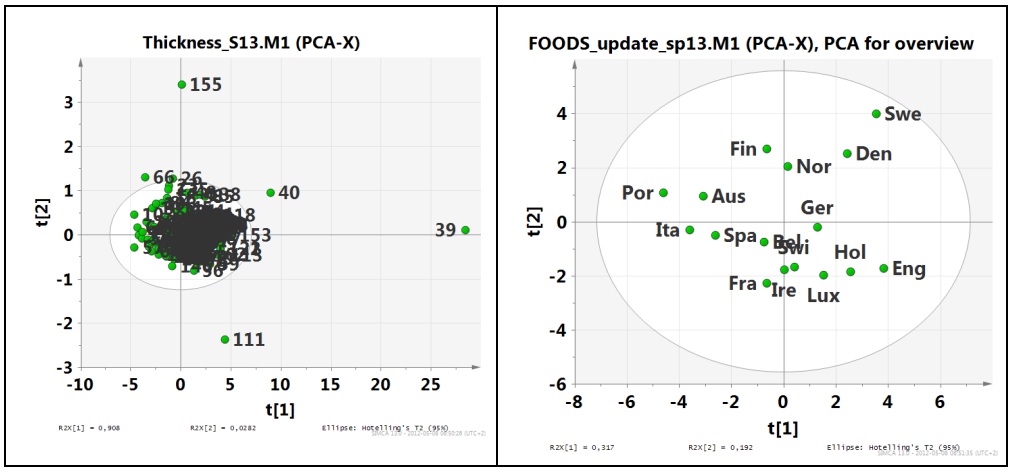


*Figure 3.24: (left) PCA score plot of a process. Observation 39 is a very strong outlier and has shifted PC1 with respect to itself. Also observations 40, 111, and 155 fall clearly outside the confidence ellipse. Figure 3.25: (right) PCA score plot of the FOODS dataset. This dataset contains no strong outlier. Recall that with 16 observations around 16*0.05 = 0.8 observation is expected to be outside the Hotelling's T^2 tolerance ellipse.*

In Figure 3.24, the extreme character of observation number 39 is beyond all doubt. This sample, and perhaps also samples 40, 111 and 155, ought to be more closely inspected. In order to better resolve and examine the main cluster, the four indicated samples must be temporarily omitted and a new PCA model fitted. Notice, however, that an outlier of this sort (cf. #39) may well be the interesting case, which should be looked at in detail in future investigations.

Figure 3.25 shows the first score plot of the FOODS dataset. Here, no strong outliers are seen. The data points are fairly evenly scattered and it is easy to overview the relationships among them. In addition, it must be stressed that with *N* observations, it is to be expected that around N*0.05 observations will be found outside the 95% confidence region. Only a handful of these "potential" outliers are likely to be "real" outliers. Hence, in any modeling, insight into the process is required to sort out the "real" ones.

A dataset may also contain moderate outliers, which are not powerful enough to shift the model plane and hence show up as outliers in score plots. Moderate outliers are identified by the residuals of each observation. In SIMCA, the detection tool for moderate outliers is called DModX, a short-hand notation for *distance to the model in X-space* (Figure 3.26). DModX is based on considering the elements of the residual matrix **E** and summarizing these row-by-row (Figure 3.27).

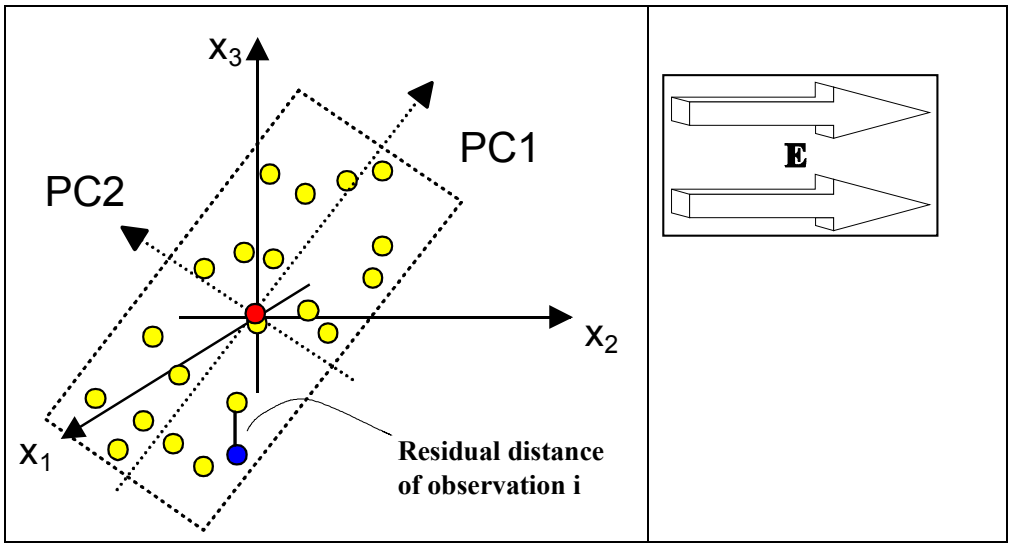


Figure 3.26: (left) A geometrical interpretation of an observation's distance to the model (DModX). A value for DModX can be calculated for each observation and these values may be plotted against the time order or observation number together with a typical deviation distance (Dcrit) in order to reveal moderate outliers. Figure 3.27: (right) In the computation of DModX, the residuals of the matrix E are summarized row-by-row.

A value for DModX can be calculated for each observation. These values can be plotted in a control chart where the maximum tolerable distance (Dcrit) for the dataset is given (cf. Figure 3.28). Moderate outliers have DModX values larger than Dcrit. When DModX is twice Dcrit, there is indication of more serious outliers. Such observations are different from the normal observations with respect to the correlation structure among the variables. With process data, moderate outliers often correspond to temporary process upsets, but occasionally more persistent trends or shifts can be diagnosed (Figure 3.29). For process diagnostics, it is of vital importance to uncover outliers in residuals, as a persistently high occurrence of outliers indicates a shift in process behavior.

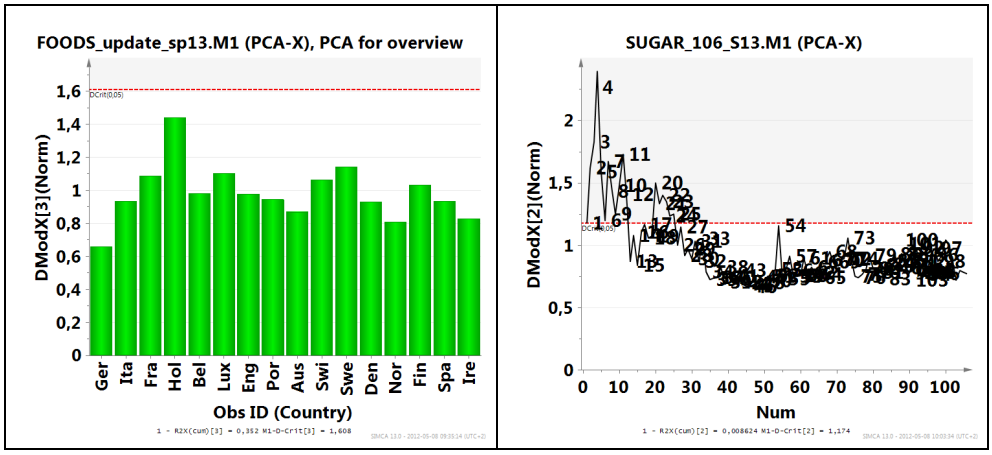


Figure 3.28: (left) DModX control chart for the FOODS dataset. No observation exceeds the critical distance (Dcrit). Figure 3.29: (right) DModX chart for a process industry, showing a typical short-term deviation (process upset) at the beginning of the sampling campaign. After a while, the process reaches a state of stability.

Using the SIMCA terminology, the calculation of DModX can be formalized as follows. The residual observation variance, S2OX, is computed as

$$S2OX = \sum_k e_{ik}^2 / DF \qquad \text{(eqn. 3.2)},$$

where DF represents the number of degrees of freedom. Here the index *i* represents the observations and the index *k* the variables. The residual observation variance can be converted to the absolute distance DModX as

$DModX_{abs} = (S2OX)^{1/2}$ (eqn. 3.3),

or the normalized distance DModX as

$DModX_{norm} = [S2OX/variance\ (E)]^{1/2}$ (eqn. 3.4).

In SIMCA, it is possible to plot S2OX, $DModX_{abs}$, and $DModX_{norm}$ in control charts. A more thorough account of these parameters is provided in Appendix II.

3.5.2 Variable diagnostics – Which variables are well-explained?

Apart from pooling the elements of the E-matrix row-wise, these elements may also be summarized column-wise to produce diagnostics related to the variables (Figure 3.30). One such diagnostic tool is called the *explained variation* of a variable, a quantity which ranges from 0 (no explanation) to 1 (complete explanation). It tells us the extent to which each variable is accounted for by the model.

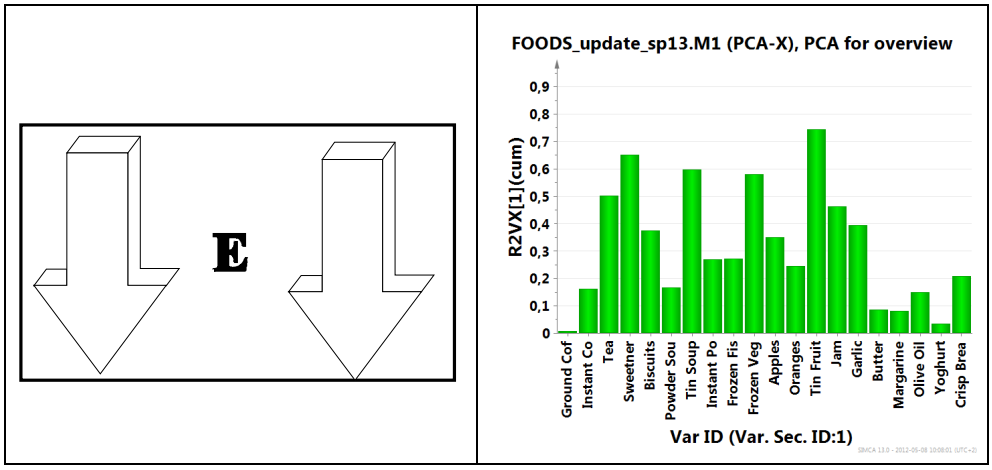


Figure 3.30: (left) In the formation of variable related diagnostics, the entries in matrix E are summarized column-by-column. Figure 3.31: (right) Explained variation of the variables of the FOODS dataset after the first PC.

Figures 3.31 – 3.33 demonstrate how the explained variation of each variable in the FOODS dataset is altered by increasing the number of principal components in the model. The limiting case is 15 (*N-1*) components because there are 16 (*N*) observations, and in that situation the explained variance of each variable is 1. The latter is an unrealistic model, however.

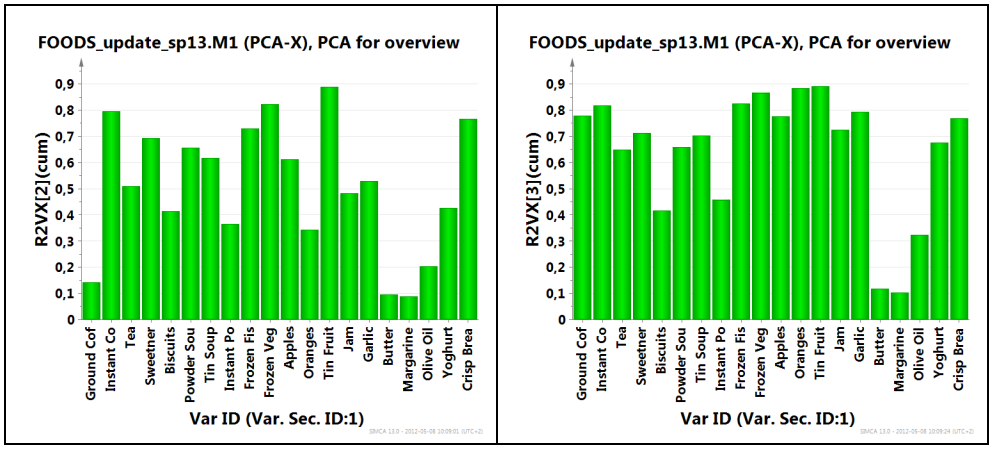


Figure 3.32: (left) Same as Figure 3.31, but after the second component. Figure 3.33: (right) Same as Figure 3.31, but after the third component. Some variables, like Butter and Margarine, are not well explained by the three-component model.

The sequence of Figures 3.31 – 3.33 makes it possible to follow how the individual variables are modeled by the different principal components. Some variables are well accounted for by the first PC, some make their entrance in the model thanks to the second PC, whereas some wait until the last PC. The variable Tin_Fruit is an example of a variable that is predominantly modeled by the first PC, and Gr_Coffe of a variable that is exclusively captured by the third PC. Then, there are variables which load onto more than one component, for instance Yoghurt, which is rather well mapped by the second and the third principal components jointly.

The calculations and the SIMCA terminology involved in the derivation of the residuals-based variable diagnostics are given below. By the column-wise summation of the residual elements of **E**, it is possible to describe how well a variable is modeled by the calculation of its explained variation (R^2X) or explained variance (R^2X_{adj}). The explained variance is simply the explained variation adjusted for the degrees of freedom (DF). The values of R_k^2 are related to the loadings. For each component, a, p_{ak}^2 is proportional to how much the k^{th} variable is modeled by this component.

The residual variable variation, *SSVX*, is computed as

$$SSVX_k = \Sigma_i e_{ik}^2 \quad \text{(eqn. 3.5).}$$

This gives the corresponding residual variable variance, $S2VX_k$, by dividing by the degrees of freedom

$$S2VX_k = \Sigma_i e_{ik}^2/DF \quad \text{(eqn. 3.6).}$$

The explained variation and the explained variance of a variable are given by

$$R2VX_k\ (cum) = 1 - SSVX_k[A]/SSVX_k[0] \quad \text{(eqn. 3.7)}$$

$$R2VXadj_k(cum) = 1 - S2VX_k[A]/S2VX_k[0] \quad \text{(eqn. 3.8),}$$

where *A* represents the number of principal components. A more thorough description of these parameters is found in Appendix II. Also observe that it is possible to calculate R^2X- and R^2X_{adj}-values pertaining to the complete X-matrix (not just to the individual variables), and this is discussed in Section 3.5.3.

3.5.3 Model diagnostics – How many principal components are really needed?

An important question is *how many components should be included in the model*? This question is linked to the difference between the degree of fit and the predictive ability. The fit tells how well we are able to mathematically reproduce the data of the training set. A quantitative measure of the *goodness of fit* is given by the parameter R^2X (= the explained variation). The problem with the goodness of fit is that with sufficiently many free parameters in the model, R^2X can be made arbitrarily close to the maximal value of one (1.0).

More important than fit, however, is the predictive ability of a model. This can be estimated by how accurately we can predict the X-data, either internally via existing data or externally through the use of an independent validation set of observations. The predictive power of a model is summarized by the *goodness of prediction* parameter Q^2X (= the predicted variation). Here, we use cross-validation (CV) to estimate the predictive ability of the model with increasing number of components (see next section).

The R^2X- and Q^2X-parameters display entirely different behavior as the model complexity increases (Figure 3.34). The goodness of fit, R^2X, varies between 0 and 1, where 1 means a perfectly fitting model and 0 no fit at all. R^2X is inflationary and approaches unity as model complexity (number of model parameters, number of components, …) increases. Hence, it is not sufficient to have a high R^2X. The goodness of prediction, Q^2X, on the other hand, is less inflationary and will not automatically come close to 1 with increasing model complexity, provided that Q^2X is correctly estimated.

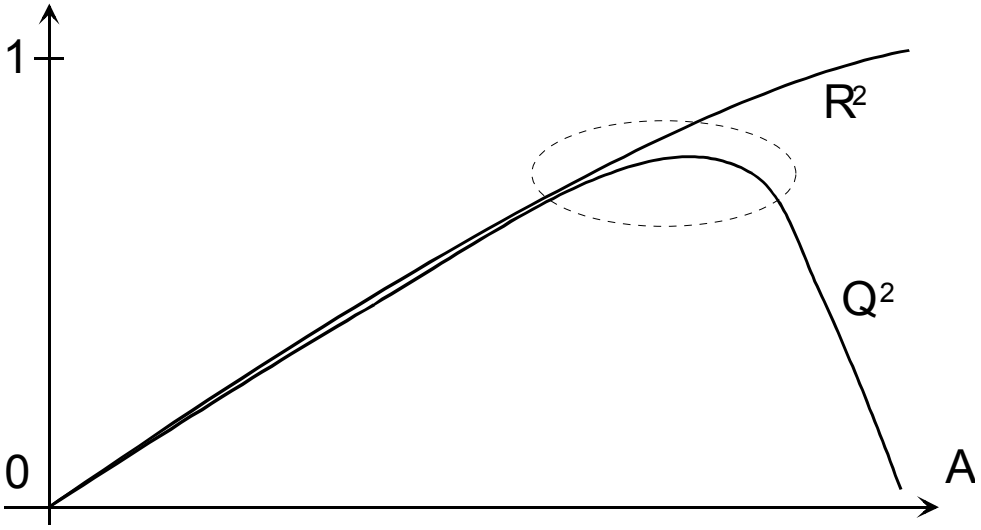


Figure 3.34: The trade-off between the goodness of fit, R^2X, and the goodness of prediction, Q^2X. The vertical axis corresponds to the amount of explained or predicted variation, and the horizontal axis depicts the model complexity (number of terms, number of latent variables, etc). At a certain model complexity, one gets the model with optimal balance between fit and predictive ability.

Hence, by a valid model we mean that it predicts much better than chance. In addition, it should have model parameters with little bias, i.e., they should have the correct sign and be large for important variables and small for unimportant variables. Finally, it should be consistent with fundamental biological, chemical, technical and engineering knowledge.

Using the LOWARP example and a PCA model of the 14 responses, we will now see how the R^2X- and Q^2X-values can be used in the evaluation of a PCA model. Figure 3.35 displays the evolution of the parameters R^2X and Q^2X with increasing model complexity. Clearly, in this case it is not meaningful to use more than three principal components, because the predictive ability does not increase as a result of the fourth and higher-order components.

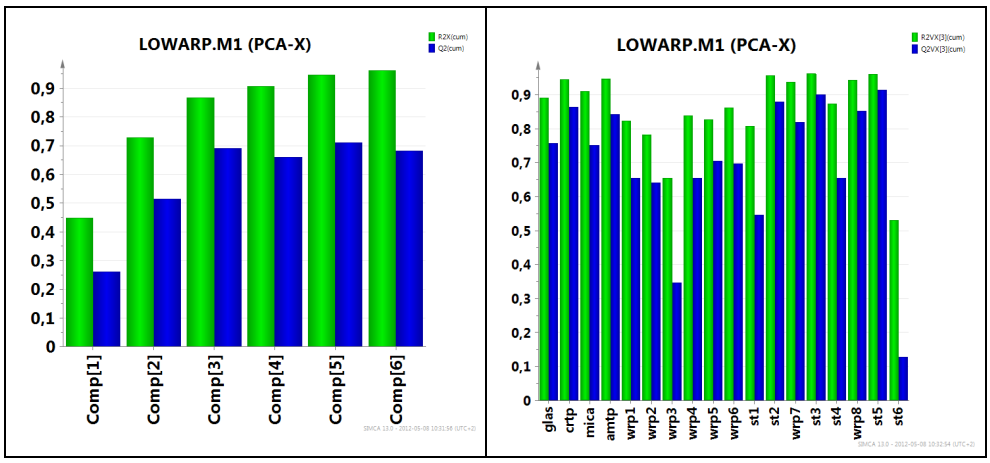


Figure 3.35: (left) Model overview of the PCA model for the 14 responses of the LOWARP dataset. The plot shows how the overall R^2X- and Q^2X-statistics change as a function of increasing model complexity. Here, three components appears appropriate, as Q^2X does not increase beyond the third component. Figure 3.36: (right) The overall R^2X- and Q^2X-values may be decomposed into R^2's and Q^2's relating to the individual variables. The situation after three model components is depicted. Apparently, most variables are rather well modeled after three PCs, but there are some exceptions. In particular, the analyst should focus on understanding what is wrong with the variables wrp3, wrp4 and st6, with large gaps between R^2X and Q^2X.

It is possible to partition the overall R^2X- and Q^2X-estimates into values pertaining to the 14 individual variables. Figure 3.36 presents these after three principal components. For the majority of the variables the computed PCA model is able to well describe their variation.

But there are some exceptions. In particular, the three variables wrp3, wrp4 and st6 have large gaps between R^2X and Q^2X. Such large gaps indicate problems with these variables.

3.5.4 Cross-validation

The approach to finding the optimal model dimensionality advocated throughout this course book is called cross-validation [Wold, 1978]. Cross-validation (CV) is a practical and reliable way to test the significance of a PCA- or a PLS model. This procedure has become standard in multivariate analysis, and is incorporated in one form or another in most commercial software. However, CV is implemented differently in different packages, which may cause some confusion when comparing models developed by different packages.

With CV the basic idea is to keep a portion of the data out of the model development, develop a number of parallel models from the reduced data, predict the omitted data by the different models, and finally compare the predicted values with the actual ones (Figure 3.37). The squared differences between predicted and observed values are summed to form the *predictive residual sum of squares* (PRESS), which is a measure of the predictive power of the tested model. PRESS is computed as

$$PRESS = \Sigma (x_{ik} - \hat{x}_{ik})^2 \qquad \text{(eqn. 3.9).}$$

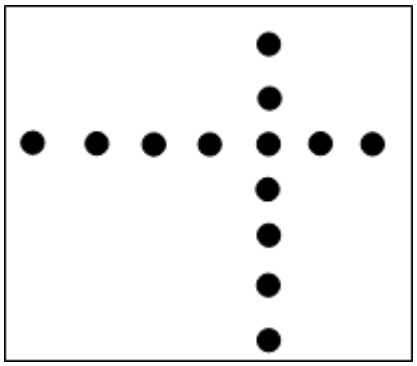


Figure 3.37: In PCA, data are divided into G groups, typically between 5 and 10 groups, and temporarily deleted in a combined row-wise/column-wise fashion. A model is established for the data devoid of one group. The deleted group of elements is predicted by the model and the predictive residual sum of squares (PRESS) is calculated. This procedure is repeated G times, followed by the summation of all partial PRESS-values in terms of an overall PRESS-value. If a new PC_a enhances the predictive power compared with the preceding PC_{a-1}, the new PC_a is kept in the model.

In SIMCA, CV is conducted for each consecutive model dimension starting with A = 0. For each additional dimension, CV gives a PRESS, which is compared with the residual sum of squares (RSS) of the previous dimension. When PRESS is not significantly smaller than RSS, the tested dimension is considered insignificant and the model building is stopped.

Normally, the performance of a PCA model in SIMCA is evaluated by simultaneously considering the explained variation R^2X (goodness of fit) and the predicted variation Q^2X (goodness of prediction). As shown by equations 3.10 and 3.11, these two statistics resemble each other:

$$R^2X = 1 - RSS/SSX_{tot.corr.} \qquad \text{(eqn. 3.10)}$$

$$Q^2X = 1 - PRESS/SSX_{tot.corr.} \qquad \text{(eqn. 3.11),}$$

and they are both dimensionless. In the expressions above, $SSX_{tot.corr.}$ represents the total variation in the X-matrix after mean-centering.

In the evaluation of the parameters R^2X and Q^2X, there are a few noteworthy facts. The first is that without a high R^2X it is impossible to get a high Q^2X. Generally, a $Q^2X > 0.5$ is regarded as good and a $Q^2X > 0.9$ as excellent, but these guidelines are of course heavily application dependent. Finally, the difference between R^2X and Q^2X must not be too large, and preferably not exceeding 0.2 - 0.3.

3.6 Questions for Chapter 3

1. Which two approaches are commonly used for data pre-processing?
2. What is a principal component?
3. What is a *score*?
4. What is a *loading*?
5. What is a strong outlier?
6. Which detection tool can be used to uncover strong outliers?
7. What is a moderate outlier?
8. Which detection tool can be used to identify moderate outliers?
9. Which statistics indicate which variables are well explained?
10. How do we determine the optimal dimensionality of a PCA model?

3.7 Summary and discussion

Principal component analysis summarizes the variation of a data matrix X, as a product of two low-dimensional matrices, T and P´, which may be easily overviewed and used. The data in the multi-dimensional space, defined by the measured variables, is modeled as a plane or hyperplane, the axes of which are called the principal components. Each principal component can be displayed graphically and may often be interpreted according to chemical, technical and/or biological knowledge.

Prior to PCA, data are typically pre-processed by mean-centering and scaling to unit variance. With PCA, the X-matrix is then decomposed ($X = 1*\bar{x} + T*P' + E$) into the product of two matrices, the (N x A) score matrix T times the (A x K) loading matrix P´, plus an (N x K) "noise" matrix of residuals, E. The value of A, the number of principal components, is usually determined by cross-validation. Alternative stopping rules for determining the number of relevant principal components is discussed by [Peres-Neto, *et al.*, 2005]. As long as the dimensionality of the projection, A, is small, this model of the variation in X as T*P´ is a considerable simplification.

Plots of PCA scores show the relationships among the observations including outliers. Plots of residuals, for example DModX, pinpoint moderate outliers. To support decision making some statistical tools are available. For score plots, Hotelling's T^2 may be used to define a 95% or 99% tolerance region. For DModX, a critical distance corresponding to the 0.05 level is available. Observations outside the ellipse defined by Hotelling's T^2 deviate from normality, and observations exceeding the critical distance in DModX do not fit the model well.

We make a distinction between strong and moderate outliers. Strong outliers, seen in score plots, *conform* with the overall correlation structure of the data (although they may have an extreme character), whereas moderate ones, which are found in residual plots, *break* this correlation structure. A moderate outlier does not have the same profound effect on the PCA model as a strong outlier does. But for diagnostic purposes it is important to identify outliers in residuals, because they indicate lack of homogeneity ("upsets") in X.

Another attractive property of PCA is that it applies to almost any type of data matrix, e.g., matrices with many variables (columns), many observations (rows), or both. The precision and reliability of the PCA parameters related to the observations (scores, DModX) is enhanced by increasing the number of relevant variables. This property is readily understood by realizing that the "new variables", the scores t_a, are estimated as weighted averages of the X-variables. Any (weighted) average becomes more precise the more numerical values are used as its basis. Hence, PCA works well with "short and wide" matrices, i.e., matrices with many more columns than rows.

Analogously, the PCA parameters supplying information related to the variables, for example, loadings, R^2X, Q^2X, etc., become more reliable the larger the number of observations. This because the loadings, etc., are linear combinations, i.e., *weighted*

averages, of the *N* observation vectors. Thus, PCA also works well with "long and lean" data structures, i.e., data tables with many more rows than columns.

Obviously, PCA is also capable of handling "square" matrices, i.e., situations in which the number of rows and columns are approximately equal.

Because of its power and versatility PCA today goes under many names. Singular value decomposition (SVD) is a common name in numerical analysis, Karhunen-Loeve expansion in electrical engineering, and Eigenvector analysis is often used in the physical sciences. In some French-speaking countries a special double-scaled variant of PCA, known as Correspondence analysis, is much favored (see Chapter 13). And, finally, we have the closely related method Factor Analysis (FA), which was originated by psychologists in the 1930s. PCA and FA are often confused and often the two names are incorrectly used interchangeably [Jackson, 1991].

PCA also forms the basis for hierarchical modeling (see Chapter 19). With the hierarchical approach, score variables, derived from multiple PCA models at the lower modeling level, are concatenated to form new upper modeling level data matrices that are again analyzed by PCA. The upper level provides a superficial overview of the relationships among the different blocks of variables. The lower level allows zooming into the most influential variables in the important variable blocks. The use of PCA as a corner-stone in data mining using hierarchical cluster analysis (HCA) is delineated in Chapter 21.

4 PLS

4.1 Objective

The objective of Chapter 4 is to describe the second main multivariate data analytical tool used in this book, namely partial least squares projections to latent structures, or PLS. PLS is a regression extension of PCA, which is used to connect the information in two blocks of variables, X and Y, to each other. As with PCA in the preceding chapter, PLS will be explained using a geometrical representation. We will describe situations where only one response variable is modeled, and where several responses are analyzed at the same time. PLS provides many model parameters and other residuals-based diagnostic tools, which are useful for understanding and interpreting the acquired regression model. These diagnostics are described and illustrated.

4.2 Introduction to PLS

4.2.1 Some terminology

In its simplest form – the form most often used in chemistry and process modeling – PLS is a method for relating two data matrices, X and Y, to each other by a linear multivariate model (Figure 4.1) [Höskuldsson, 1996; Wold, *et al.*, 1984; Wold, *et al.*, 1999; Wold, *et al.*, 2000; Wold, *et al.*, 2001b]. PLS stands for *projections to latent structures by means of partial least squares*. It derives its usefulness from its ability to analyze data with many, noisy, collinear, and even incomplete variables in both X and Y. For parameters related to the observations (samples, compounds, objects, items), the precision of a PLS model improves with the increasing number of relevant X-variables. This corresponds to the intuition of most chemists, technicians, and engineers that many variables provide more information about the observations than just a few variables do.

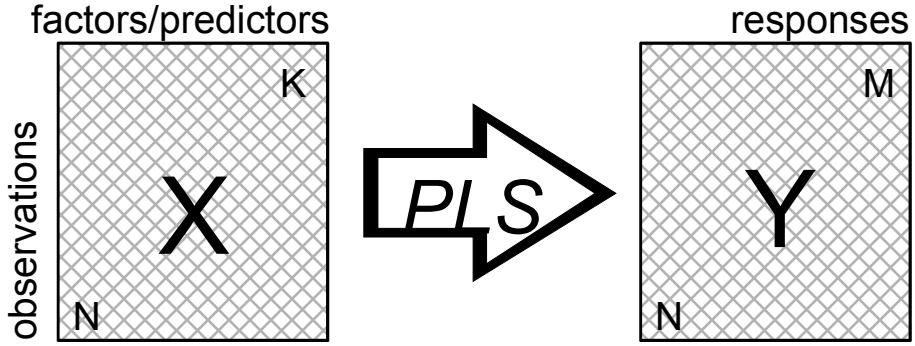


Figure 4.1: Some terminology often used in conjunction with regression and PLS. The observations can be samples from analytical chemistry, chemical compounds, process time points of a continuous process, batches from a batch process, biological individuals, runs of an experimental design, etc. The X-variables, also known as factors or predictors, can be, for instance, of spectroscopic or chromatographic origin, or be readings from sensors and devices in a process. The Y-variables, also called responses, are often gathered to reflect properties of samples or yields, impurities, and product quality of a process.

PLS can be seen as a particular regression technique for modeling the association between X and Y, but it can also be seen as a philosophy of how to deal with complicated and approximate relationships. In particular, PLS has properties similar to thermodynamic

models, in that models linking macroscopic variables are formed without recourse to microscopic theory. This does not mean that PLS is inconsistent with microscopic and detailed theory, but rather that PLS uses the regularities in the data themselves as building blocks in the models, instead of hypothetical functional forms derived from an underlying theory. For this reason, projection models like PLS also work in cases where no good fundamental theory exists, and also provide an independent check on the validity of models derived from theory, where these do exist.

4.2.2 Common application areas

Historically, the PLS methodology has proven successful in three application areas. These areas are (i) quantitative structure-activity relationships (QSAR) modeling, (ii) multivariate calibration, and (iii) process monitoring and optimization.

In QSAR, PLS is used to model the relationship between, on the one hand, computed and measured variables that together characterize the structural variation in a set of N compounds, and, on the other hand, biological response variables or other interesting properties measured on the same N substances. The former comprise the matrix X of predictor variables, and the latter the matrix Y of response variables. These matrices have dimensions (N x K) and (N x M) respectively. We note that the X-variables (predictors, factors,…) are usually not independent of each other.

In analytical chemistry PLS is mainly used for multivariate calibration. Here X contains N spectra digitized at K wavelengths, and Y the analyte concentrations of the N training set samples. The resulting model can then be used to predict the analyte concentrations from the spectra of new samples. Since the whole spectra are used in the modeling, the multivariate PLS model results in better predictive precision, and also much improved selectivity in comparison with traditional univariate calibration.

In process modeling PLS is used to find relationships between variables measured on the process (X) at N time points and corresponding values of "result variables" (Y), such as, yield, purity, and product properties, e.g., strength, viscosity and color of a polymer, taste of whisky, or kappa number of pulp.

4.2.3 A historical note

The PLS approach was originated around 1975 by Herman Wold for the modeling of complicated datasets in terms of chains of matrices (blocks), so called path models [Höskuldsson, 2001; Tenenhaus, *et al.*, 2005]. Herman Wold developed a simple but efficient way to estimate the parameters in these models called NIPALS (Nonlinear Iterative Partial Least Squares). This led, in turn, to the acronym PLS for these models, where PLS stood for Partial Least Squares. This term describes the central part of the estimation, namely that each model parameter is iteratively estimated as the slope of a simple bivariate regression (least squares) between a matrix column or row as the y-variable, and another parameter vector as the x-variable. So, for instance, in each iteration the PLS weights, w, are re-estimated as u'X/(u'u). (Here u' denotes u-transpose, i.e. the transpose of the current u-vector. PLS weights and other model parameters are described in Section 4.5.7.) The "partial" in PLS indicates that this is a partial regression, since the parameter vector (u in the example above) is considered as fixed in the estimation. This partial least squares interpretation shows that we can see any matrix-vector multiplication as equivalent to a set of simple bivariate regressions. This provides an intriguing connection between two central operations in matrix algebra and statistics, as well as giving a simple way to deal with missing data.

Around 1980 the simplest PLS model with two blocks (X and Y) was slightly modified by Svante Wold and Harald Martens to better suit to data from science and technology, and shown to be useful for dealing with complicated datasets where ordinary regression was difficult or impossible to apply. To give PLS a more descriptive meaning, S. Wold *et al.* have also recently started to interpret PLS as Projections to Latent Structures.

4.2.4 Theoretical foundation

PLS and other projection methods have a theoretical foundation based on perturbation theory of a multivariable system. This derivation shows that projection models can approximate any data table as long as there is a certain degree of similarity between the observations (matrix rows). And the approximation is better the greater the similarity between the observations and the greater the number of model components. This is very similar to the derivation of polynomials as Taylor series that can approximate any continuous function in a limited interval (see Section 2.7).

4.3 Pre-processing of data

As in any data analytical application, data are usually pre-processed prior to using PLS. PLS modeling works best when the data are fairly symmetrically distributed and have a fairly constant "error variance". Hence, variables that vary more than a ten-fold are often logarithmically transformed before the analysis. Such transformations are addressed in Chapter 12.

In addition, data are usually centered and scaled to unit variance before the analysis. This is because in PLS any given variable will have an influence on the model parameters which increases with the variance of the variable. Scaling all variables to unit variance corresponds to the assumption that all variables are equally important. Scaling is discussed more fully in Chapter 13.

If *a priori* knowledge about the relative importance of variables (X or Y) is available, this should be used to scale the variables accordingly, giving important variables a slightly higher scaling weight than that corresponding to unit variance scaling; analogously, unimportant variables are given a slightly lower scaling weight.

In multivariate calibration, additional pre-processing with so called signal correction is common to remove variation in X that is due to light scattering and other sources unrelated to Y (see Chapters 8 and 14).

Mean-centering and scaling to unit variance are the default pre-processing operations in SIMCA. A graphical illustration of the impact of these two operations is given in Figures 4.2 and 4.3.

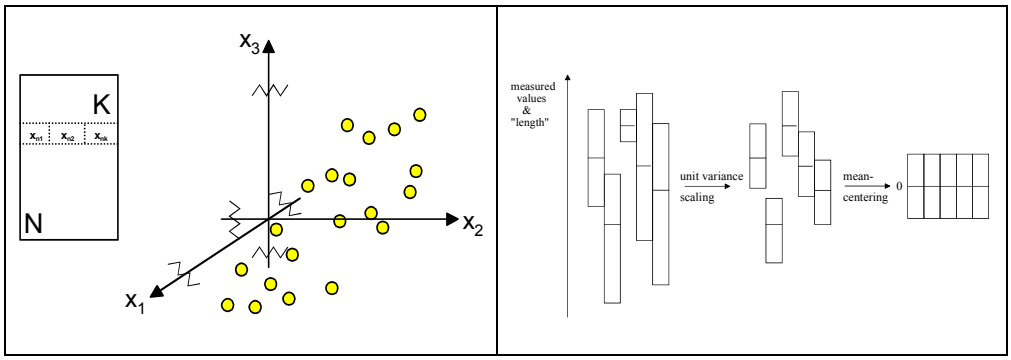


Figure 4.2: (left) The scaling of variables (both predictors and responses) means that the length of each co-ordinate axis in a variable space is regulated according to some specified criterion. The most common criterion is that the length of each variable axis be the same. In the drawing, the wavy lines denote the "start" and "end" of the variables. Figure 4.3: (right) After mean-centering and unit variance scaling all variables will have equal "length" and mean value zero.

4.4 The geometry of PLS in the case of one response (M = 1)

In this section, the basics of PLS are explained in geometric terms. The geometry of PLS has been explored in depth by Phatak and DeJong [Phatak and DeJong, 1997]. By necessity, this section will closely resemble the introduction to PCA provided in Section 3.4.

Whenever appropriate, reference will be made to this PCA section. A more formal account of PLS is found in Appendix II.

4.4.1 Constructing the X and Y co-ordinate systems

Consider a regression application with N observations, $K = 3$ X-variables, and $M = 1$ y-variable (here, we write lower case y to denote that a single variable is considered). Then, the two variable spaces can be represented as shown in Figure 4.4.

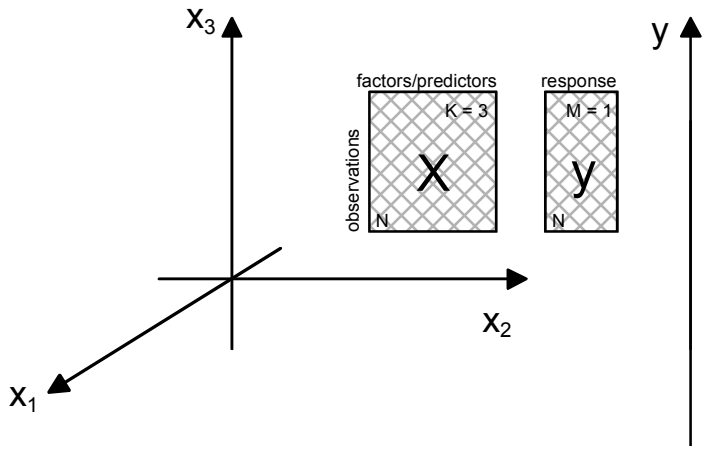


Figure 4.4: A regression situation with K = 3 X-variables and M = 1 y-variable. The length of each co-ordinate axis has been standardized by scaling to unit variance.

As in PCA, each observation can be represented graphically. However, the big difference in PLS is that each row of a data table corresponds to *two* points rather than one, one in the X-space and one in the Y-space. Consequently, with many observations in the dataset, two point-swarms are situated in these spaces. Figure 4.5 illustrates how it may look when $K = 3$ and $M = 1$. The task for the data analysis is to describe the relationship between the positions of the observations in the predictor space (X) and their positions in the response space (Y).

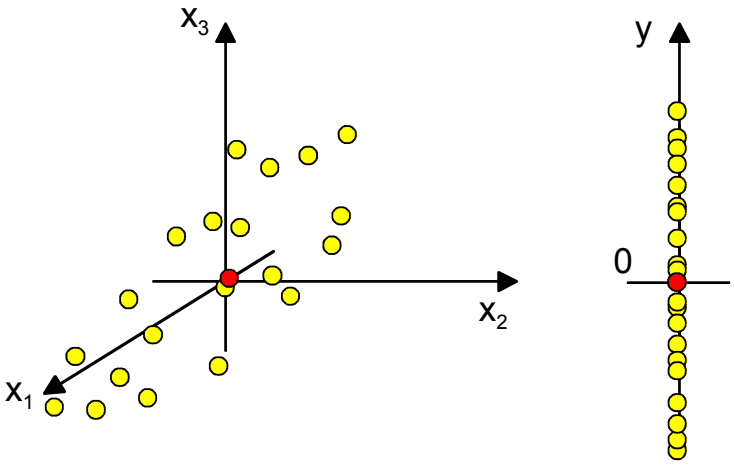


Figure 4.5: In a regression problem, the observations can be understood as two point-swarms, one in the predictor (X) space and one in the response (Y) space. Note that in the drawing above, we are considering a single variable y, rather than a matrix (Y) of responses. Data have been mean-centered. Thus the co-ordinate system in X and the co-ordinate vector in y pass through the average point (in red) of each point-swarm.

4.4.2 The first PLS component (M = 1)

After mean-centering and scaling to unit variance, it is time to calculate the first PLS component. This component is a line in the X-space which well approximates the point-swarm and provides a good correlation with the y-vector (Figure 4.6). The co-ordinate of an observation along this line is obtained by projecting the sample onto the line (Figure 4.6).

This co-ordinate is termed the *score*, t_{i1}, of observation *i*. The scores of all the observations form the first X-score vector $\mathbf{t_1}$.

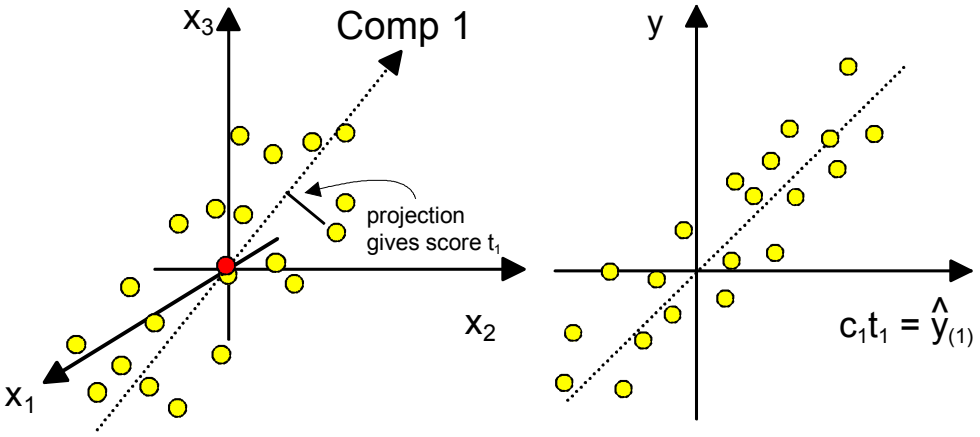


Figure 4.6: With one single y-variable, the Y-space reduces to a one-dimensional vector. The first component of the PLS model will therefore orient itself so that it well describes the point-swarm in the X-space while at the same time giving a good correlation with the y-vector. The projections of the observations onto the line in the X-space give the score of each observation. These are the new co-ordinate values of the observations and they form the score vector t_1. This vector summarizes the information in the original X-variables. A model estimate of y, $\hat{y}_{(1)}$, after the first PLS component, is accomplished by multiplying t_1 by the weight of the y-vector, c_1.

The score vector $\mathbf{t_1}$ can be thought of as a new variable, a latent variable, which reflects the information in the original X-variables that is of relevance for modeling and predicting the response variable. Subsequently, this score can be used to acquire an estimate of $\mathbf{y}$, $\hat{\mathbf{y}}_{(1)}$, after the first PLS component, which is $\mathbf{t_1}$ multiplied by the weight of the y-vector, $\mathbf{c_1}$ (see explanation in Section 4.5.7). This is illustrated in the right-hand part of Figure 4.6.

The differences between measured and estimated response data are called *residuals*. The y-residuals represent the variation that is left unexplained by the first PLS component. A good model has small residuals. The scatter of the points around the diagonal in Figure 4.6 (right-hand part) is a graphical way to assess the model performance. When all points are situated on the diagonal, we have an ideal (though not very realistic) model of the response data with zero residuals.

Another way to look at the residuals of the response vector is presented in Figure 4.7, where the residual vector $\mathbf{f_1}$, derived as $\mathbf{y} - \hat{\mathbf{y}}_{(1)}$, is plotted in the right-hand part. This residual vector is much shorter than the parent vector of observed y-values. This implies that the first PLS component has accounted for a large part of the variation in $\mathbf{y}$.

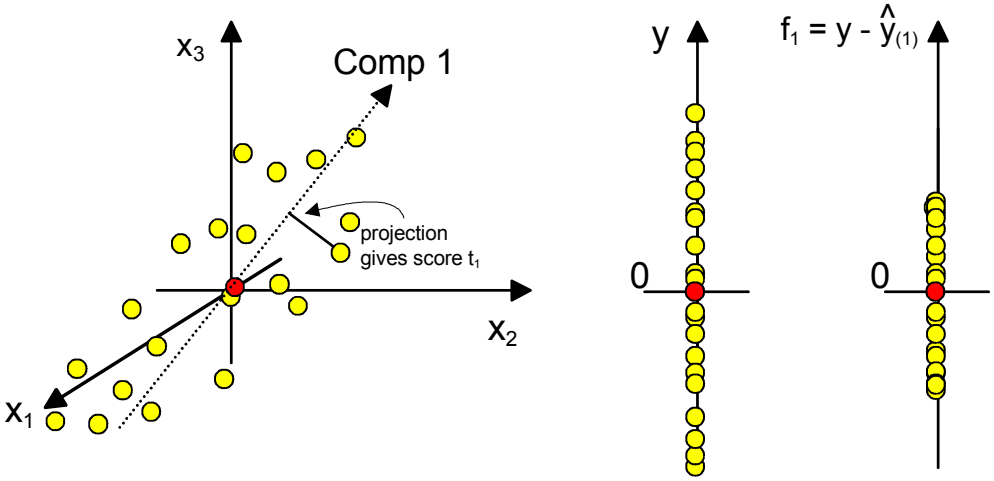


Figure 4.7: An illustration of the y-residuals remaining after the first PLS component. The residual vector f_1, obtained by subtracting $\hat{y}_{(1)}$ from y, is shorter than the vector consisting of the measured data. Therefore, the first component has explained a large part of the response variation.

4.4.3 Extending the model with the second component (M = 1)

Usually, one PLS component is insufficient to adequately model the variation in the y-data. The descriptive ability of the PLS model is then improved by expanding it with a second component. The second component is also a line in X-space, which passes through the origin and is orthogonal to the first component (Figure 4.8). This component finds the direction in X-space that improves the description of the X-data as much as possible, while providing a good correlation with the y-residuals, $\mathbf{f_1}$, remaining after the first component.

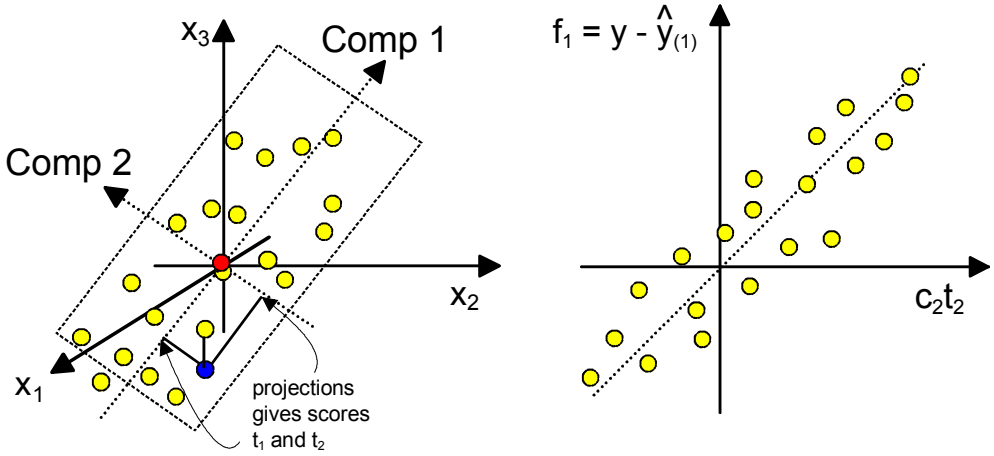


Figure 4.8: The second projection co-ordinate (i.e., the second latent variable) in the X-space is orthogonal to the first one. By projecting the observations onto this line one obtains the score vector t_2. In the right-hand part of the figure, we can see how the second score vector times the second weight of the y-data, c_2, correlates with the y-residual, f_1, after the first dimension. The two components together define a plane in the X-space.

As shown in the left-hand part of Figure 4.8, the second set of score values of the observations arises from the co-ordinates along the second projection direction in the X-space. This second score vector is denoted $\mathbf{t_2}$. In the right-hand part of Figure 4.8, we can see how $\mathbf{t_2}$ times the second weight of the y-data ($\mathbf{c_2}$, see discussion in Section 4.5.7) correlates with the y-residual, $\mathbf{f_1}$, after the first dimension. The interpretation of the right-hand part of Figure 4.8 is similar to that of the right-hand part of Figure 4.6. Thus, the tighter the scatter around the dotted diagonal, the stronger the correlation between **X** and **y** in the second PLS model dimension.

So far we have only investigated the correlation between the score vectors $\mathbf{t_1}$ and $\mathbf{t_2}$ and the y-data for one model component at a time (Figures 4.6 and 4.8). However, we can also examine the combined power of $\mathbf{t_1}$ and $\mathbf{t_2}$ in modeling and predicting **y**. An estimate of y after two model components, $\hat{\mathbf{y}}_{(2)}$, is obtained through computing $\mathbf{c_1t_1} + \mathbf{c_2t_2}$ (Figure 4.9). By comparing the right-hand parts of Figures 4.6 and 4.9 it is realized that the y-variable is better modeled by two components than by one. This is because the agreement between observed and estimated y-data is better with two components. Geometrically, this more precise estimate of **y**, $\hat{\mathbf{y}}_{(2)}$, is interpretable as the vector addition of component 1 to component 2 in the X-space (Figure 4.9).

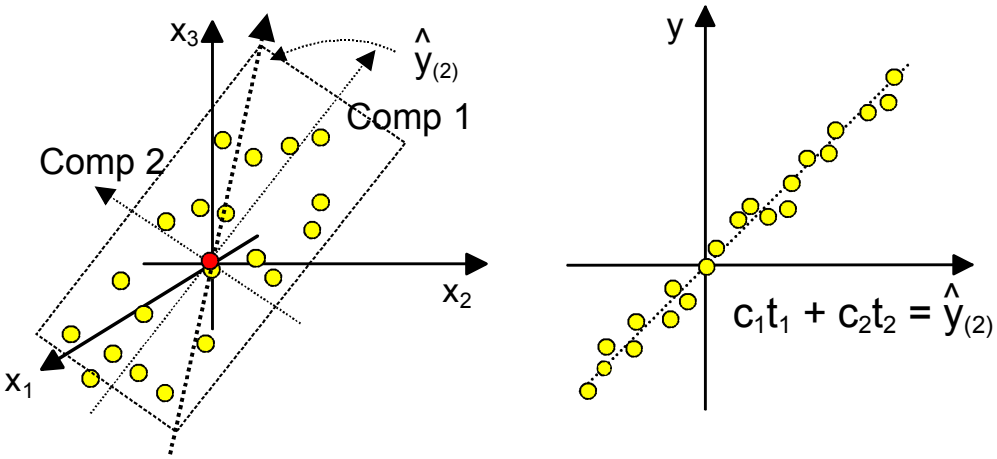


Figure 4.9: An estimate of y after two model components, $\hat{y}_{(2)}$, is obtained as $c_1t_1 + c_2t_2$. When the agreement between observed and estimated y-data is good the scatter of points around the dotted line (in the right-hand part of the figure) is small. A comparison of Figures 4.6 and 4.9 indicates that the y-variable is better modeled with two components than when using only one component. Geometrically, $\hat{y}_{(2)}$ is interpretable as the result of the vector addition of component 1 and component 2 in the X-space.

The impact of the second model dimension is illustrated in Figure 4.10. The first component has explained much of the y-variation, and only a fraction remains in the residual $\mathbf{f_1}$. The situation is further improved after the inclusion of the second component, as the residual $\mathbf{f_2}$ is smaller than $\mathbf{f_1}$ (Figure 4.10). It is easily realized that the limiting case is a model with no residuals at all, i.e., a model for which $R^2(Y)$ is unity (1.0). Such a "perfect" model is, however, usually a greatly overfitted model.

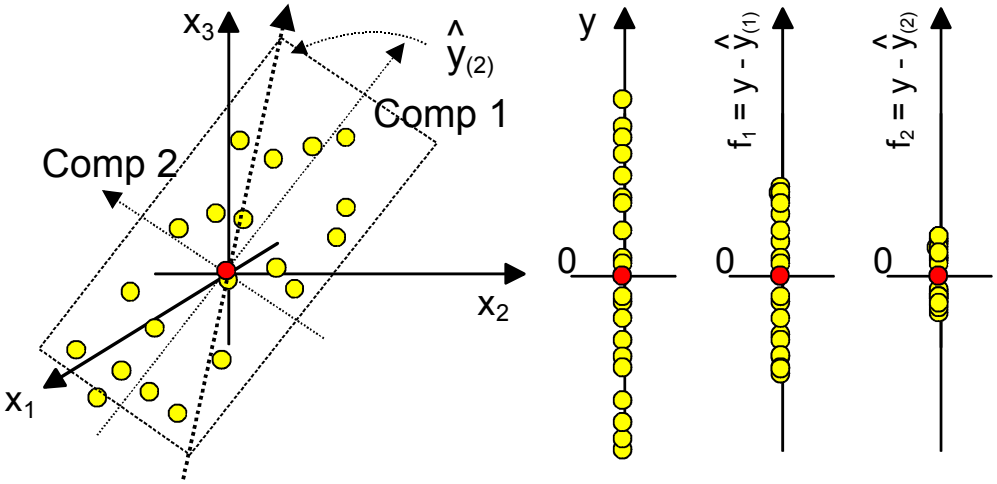


Figure 4.10: Schematic illustration of the explanatory power of a PLS model. After one component, the y-residual (f_1) is significantly smaller than the spread of the measured variable. The situation is even better after the inclusion of the second component, and the residual f_2 is smaller than f_1.

4.4.4 Example: Sugar

In the SUGAR dataset (Section 1.7.2) the objective was to use fluorescence data from sugar production (Figure 4.11) to model and predict the ash content. The spectral data were only mean-centered, i.e., no additional scaling was used.

The first preliminary analysis with PCA revealed that 16 observations (time points) represented uninteresting process conditions, a start-up period (Sections 1.7.2 and 8.4.2). Accordingly, these 16 samples (time points) were discarded and PLS was then used to relate the spectral data to the ash content.

In order not to overfit the calibration model, the data were divided in two parts, odd- and even-numbered samples. Subsequently, one part of the data was used for model training and the other part was used for model testing (validation), and vice versa. In this section, we will briefly summarize some modeling results obtained for the odd-numbered training set. Details are given in Sections 8.4.2 and 8.4.3.

Figure 4.12 shows the t_1/t_2 score plot for the PLS model. This plot can be seen as a window into spectral (X-) space, and it is conceptually similar to the graphs shown in the left-hand parts of Figures 4.8 – 4.10. Apparently, although the most extreme observations were omitted from the PLS modeling, some deviating time points are still present. However, these do not degrade the predictive ability of the model too much (see further discussion in Chapter 8).

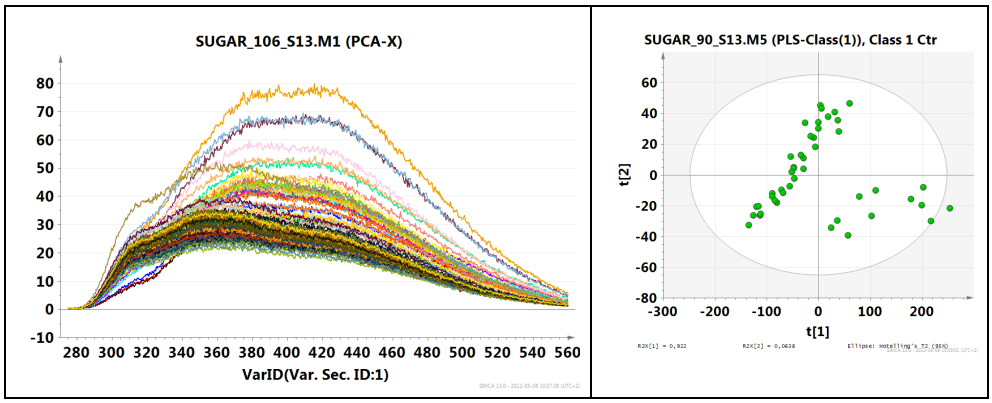


Figure 4.11: (left) Plot of raw fluorescence data. Each spectrum, when digitized, corresponds to one row of the X-matrix. These data are used in the modeling and prediction of the ash content. Figure 4.12: (right) PLS t_1/t_2 score plot of the model based on the odd-numbered training set.

The PLS model's descriptive ability of the ash content changes with the addition of more components (Figures 4.13 and 4.14). Figure 4.13 is a real example of Figure 4.6 and Figure 4.14 represents a real illustration of Figure 4.9.

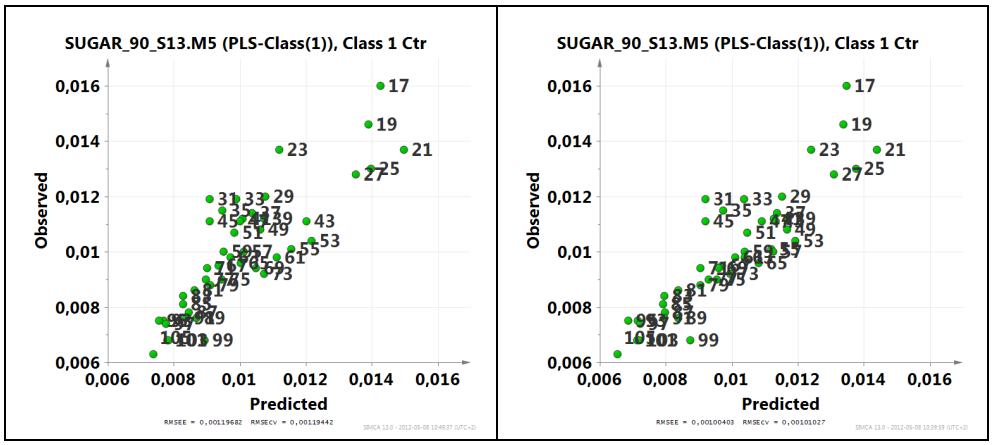


Figure 4.13: (left) Relationship between observed and calculated ash content after one PLS component.
Figure 4.14: (right) Relationship between observed and calculated ash content after two PLS components.

The differences in Figure 4.13 between observed and calculated ash content comprise the residual vector $\mathbf{f}_1$, which is plotted in a schematic manner in Figure 4.10. Likewise, the differences between observed and calculated data in Figure 4.14 constitute the residual vector $\mathbf{f}_2$, also given in Figure 4.10. Since the correlation band in Figure 4.14 is tighter than in Figure 4.13, we understand that the descriptive ability of the ash content is better with two components. Accordingly, the residual vector $\mathbf{f}_2$ has shrunk in relation to the residual vector $\mathbf{f}_1$, as is indicated in Figure 4.10. The normalized length (standard deviation, SD) of $\mathbf{f}_1$ is 0.0012, while of $\mathbf{f}_2$ it is 0.0010, i.e., a decrease of 17%.

This concludes our treatment of the geometrical representation of PLS in the case of only one response.

4.5 The geometry of PLS with multiple responses (M > 1)

4.5.1 Setting up the X- and Y-spaces

In order to illustrate how PLS operates when the response data comprise a matrix **Y** rather than a vector **y**, we will consider a case in which there are three X-variables ($K = 3$) and three Y-variables ($M = 3$). For each matrix, **X** and **Y**, we construct a space with K and M dimensions, respectively (Figure 4.15). In these two spaces, each X- and Y-variable represents a co-ordinate axis with a length defined by its scaling, usually unit variance.

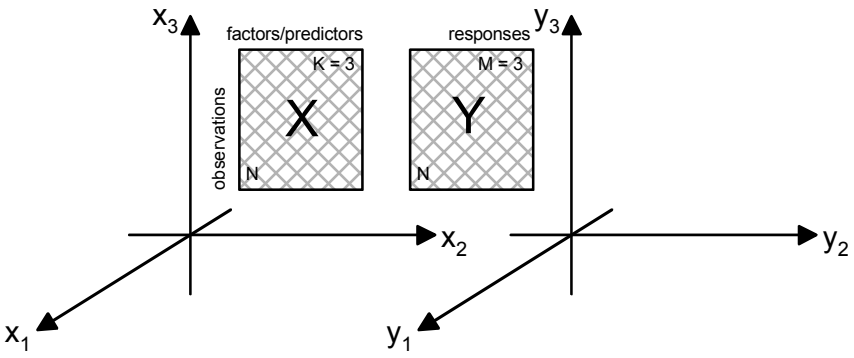


Figure 4.15: A regression situation with K = 3 X-variables and M = 3 Y-variables. The length of each co-ordinate axis has been standardized by scaling to unit variance.

Every observation in a dataset may be understood as one point in the X-space and another point in the Y-space. Thus, with many observations, point-swarms with many members are formed in the X- and Y-spaces, as illustrated in Figure 4.16. Figure 4.16 displays the mean-centered data, and the origins of the two co-ordinate systems coincide with the average points of the data swarms.

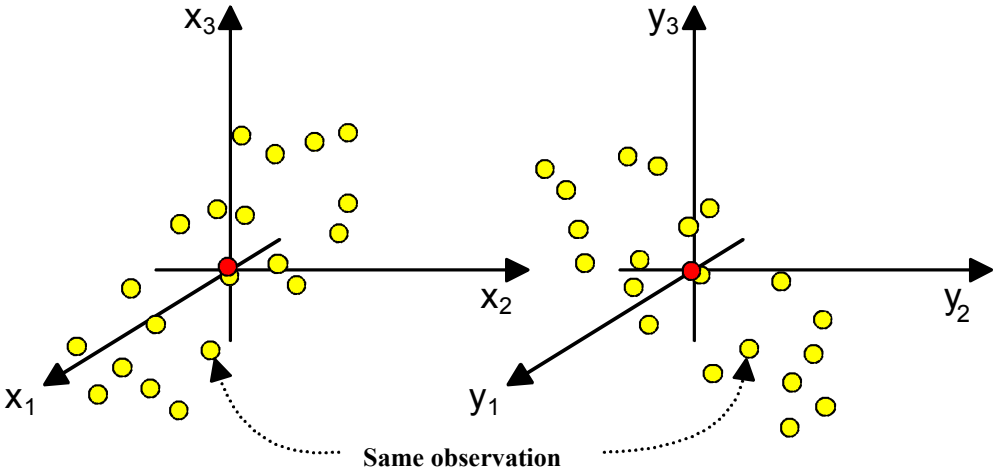


Figure 4.16: The mean-centering procedure implies that the origins of the two co-ordinate systems will coincide with the average point (red) in each cloud of points. Each observation is represented by one point in the X-space and another point in the Y-space.

The two point-swarms have elongated shapes indicating the correlated distribution of points in each cluster. We wish to obtain a good description of these two point-swarms, and an understanding of the association between them. In other words, we would like to know whether there exists a relationship between the positioning of points in the predictor space and the positioning of points in the response space. This can be elucidated by PLS analysis.

4.5.2 Calculating the first PLS component (M > 1)

The first PLS component is a line in the X-space and another line in the Y-space (Figure 4.17). These two lines are calculated such that they (i) well approximate the point-swarms in **X** and **Y**, and (ii) provide a good correlation between the positions of points along these lines in **X** and **Y**. The two lines intersect with the average points. By projecting the observations onto them, one obtains the scores $\mathbf{t_1}$ and $\mathbf{u_1}$, for **X** and **Y**, respectively (Figure 4.17).

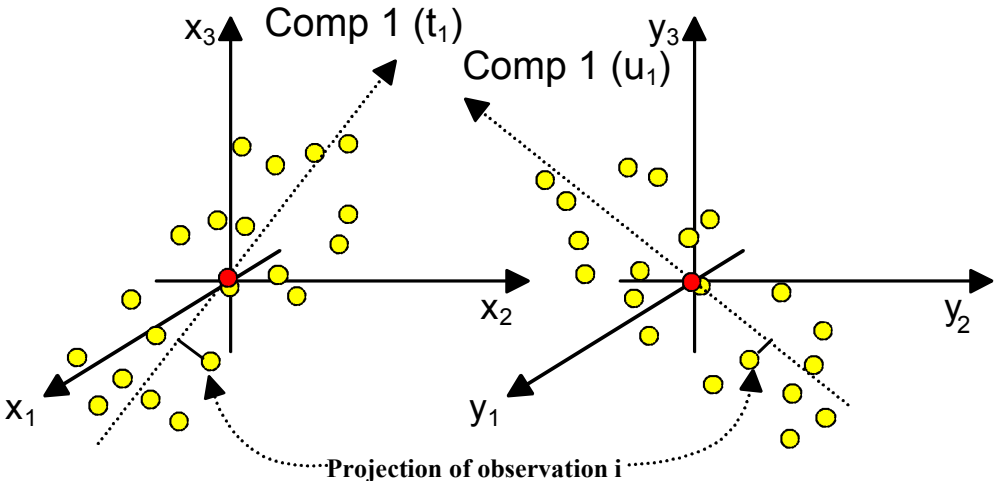


Figure 4.17: The first component of a PLS model may be interpreted as two lines, one inserted in the X-space and the other in the Y-space. The orientation of these lines is regulated by the requirement that (i) they should well approximate the shapes of the two point-swarms, and (ii) the scores t_1 and u_1 be maximally correlated. The observations projected onto the two lines give the projection co-ordinates (the "scores") t_1 (for X) and u_1 (for Y).

The correlation between **X** and **Y**, in terms of the two score vectors $\mathbf{t_1}$ and $\mathbf{u_1}$, may be displayed in a scatter plot. The two score vectors are connected through the *inner relation* $\mathbf{u_{i1} = t_{i1} + h_i}$, where $\mathbf{h_i}$ is a residual (Figure 4.18). The slope of the dotted line in Figure 4.18 is 1.0, and when there is perfect matching between the X- and the Y-data all the points are located on this diagonal. Conversely, when there is a weak correlation structure between X and Y, there is a considerable spread of points around the dotted line.

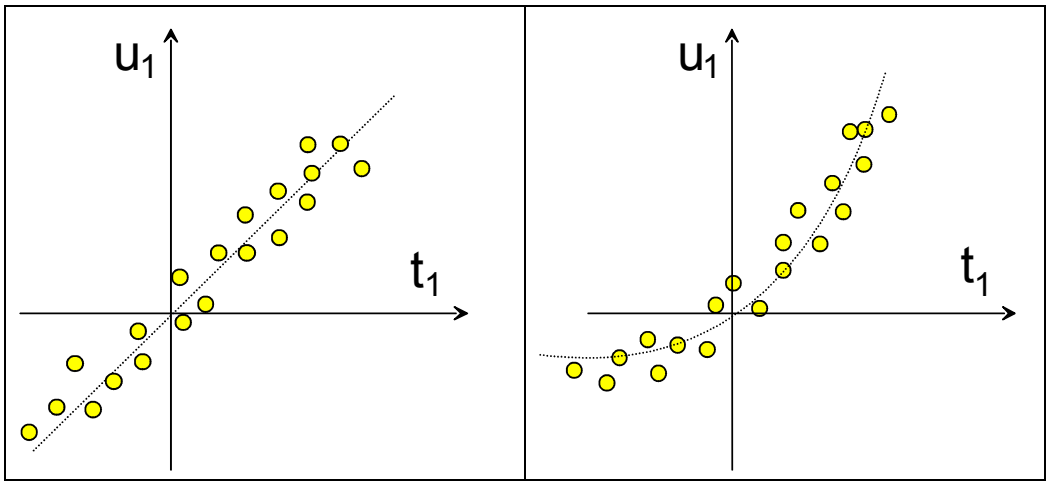


Figure 4.18: (left) The projection co-ordinates, t_1 and u_1, in the two spaces, X and Y, are connected and correlated through the inner relation $u_{i1} = t_{i1} + h_i$ (h_i is a residual). The slope of the dotted line is one (1). Figure 4.19: (right) The PLS score plot t_1/u_1 is useful for identifying curved (non-linear) relationships between the predictors and the responses.

The $\mathbf{t_1/u_1}$ score plot in Figure 4.18 is a visualization of the correlation structure between **X** and **Y.** In this score plot one can see outliers in the X-data, outliers in the Y-data, and

outliers in the relation between **X** and **Y**. Furthermore, when there are non-linearities between the predictors and the responses, these may also be detected by the slope of the relation between t_1 and u_1 (Figure 4.19 and Section 4.7.1).

4.5.3 Adding the second component (M > 1)

The second PLS component may also be represented by two lines, one in each space, which pass through the average points (Figure 4.20). In the X-space, this second line is orthogonal to the first one, whereas in the Y-space this may not necessarily be the case. These lines improve the approximation of, and correlation between, the positions of the X- and Y-planes as much as possible.

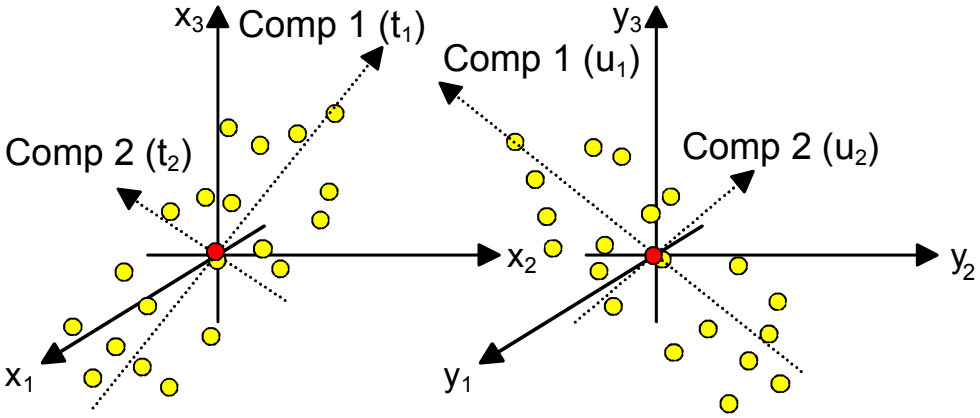


Figure 4.20: The second PLS component can be represented by two lines, one in each variable space. These lines improve the description and correlation of X and Y as much as possible.

Geometrically, a two-component PLS model can be interpreted as planes in the X- and Y-spaces. By projecting the observations onto these planes, the PLS scores $\mathbf{t_1}$ and $\mathbf{t_2}$ in **X**, and $\mathbf{u_1}$ and $\mathbf{u_2}$ in **Y**, are obtained (Figure 4.21).

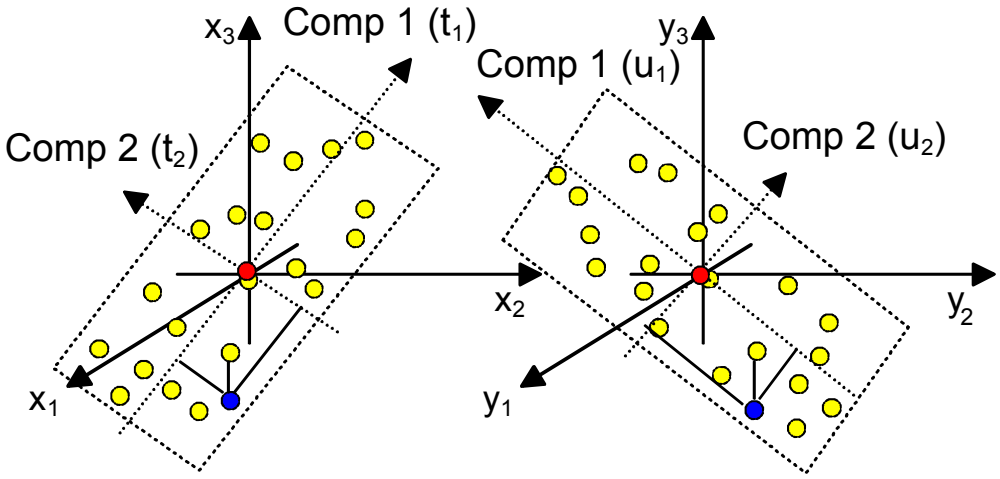


Figure 4.21: Two PLS components correspond to the insertion of model planes in the X- and Y-spaces. Upon projecting the observations onto these planes, the PLS score vectors of the first model dimension, t_1 and u_1, and the second model dimension, t_2 and u_2, are generated.

Analogously to the first score vector pair ($\mathbf{t_1}$/$\mathbf{u_1}$), a plot of the second set of score vectors, $\mathbf{t_2}$ and $\mathbf{u_2}$, also visualizes the correlation structure (Figure 4.22). Normally, the score vectors of the second component correlate less well than the first pair of latent variables. In fact, this is logical, as the first PLS component captures the strongest source of variation in the data, i.e., the strongest "signal". After the removal of the variation accounted for by the first component, weaker "signals" remain in the data, and therefore the correlation between **X** and **Y** (in terms of $\mathbf{t_2}$ and $\mathbf{u_2}$) is usually weaker and less distinct.

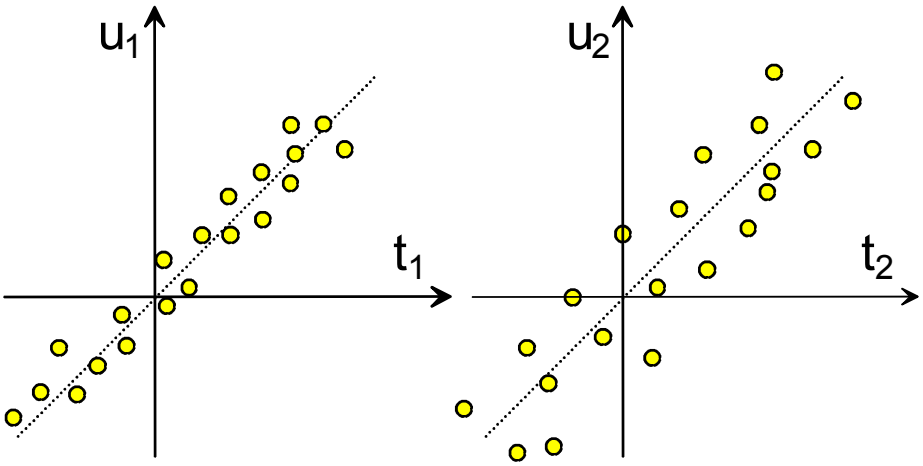


Figure 4.22: The second pair of score vectors, t_2 and u_2, are correlated, but usually less well than the first pair of score vectors (t_1 and u_1). This is indicated by the broader "correlation band" around the second component.

4.5.4 Which original variables contribute to the formation of the model planes?

Once a PLS model has been established, it is of interest to make an interpretation of its meaning. This may be accomplished by considering the variable-related PLS model parameters called *weights*. The weights for the X- and the Y-variables, which are denoted **w*** and **c**, respectively, may be plotted together in the same plot. These weights are interpreted in much the same way as the PCA loadings (see Section 3.4.7), and show which variables contribute to the PLS model, and which are not modeled at all. In principle, this means that the PLS weights reflect the relationships among all variables at the same time, and tell which are associated and which contribute unique information. Thus, with PLS, one obtains information on what **X** gives **Y**, or, how to "set" **X** to get a desired **Y**. This implies that in process modeling it might be possible not only to understand the mechanism of the process, but also to extract clues about how to modify the regulation pattern of the process to get enhanced product quality and production economy.

A technical, matrix related, description of the PLS weights is given in Section 4.5.7, and an illustration in Section 4.5.6.

4.5.5 Higher-order components (M > 1)

It is possible to include more than two components in a PLS model. When this is done, we are no longer fitting two-dimensional planes in the X- and Y-spaces, but rather hyper-planes of three, four, … , dimensions. Conceptually, such hyper-planes are no different from uni-dimensional lines or two-dimensional planes, and the principles of projecting observations onto these hyper-planes and reading off the new co-ordinate values (the scores) are preserved.

One may ask *how many PLS components are really necessary?* In Section 4.7.3 we will see how cross-validation can be utilized to shed some light on this question. Plots of successive pairs of latent variables is another way to address this topic. Not only will such plots give a good appreciation of the correlation structure, but they will also aid in determining the appropriate model complexity (Figure 4.23). Another diagnostic tool that can be used for the same purpose is the response permutation test, which is outlined in Appendix I.

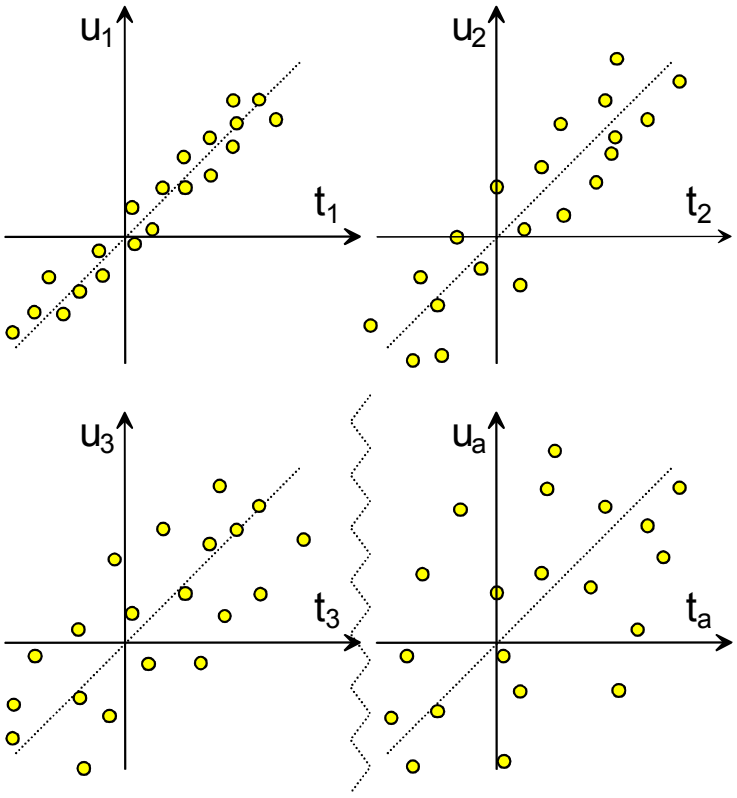


Figure 4.23: Plotting of each successive pair of score vectors (t_a/u_a) is helpful for deciding how many PLS components are relevant in a given application.

4.5.6 Example: LOWARP

The LOWARP dataset was presented in Section 1.7.1 and concerns the development of a polymer similar to that of the plastic covering of mobile phones. The desired profile of the polymer was low warp and high strength. Four constituents (*glas*, *crtp*, *mica*, and *amtp*) were varied in the polymer formulation by means of a 17 run mixture design. For each new formulation, i.e., each new experiment in the mixture design, 14 responses relating to both warp and strength were measured on the product.

The PLS analysis yielded a three-component model, but the last component is a borderline case according to cross-validation (see Section 4.7.3); therefore, in this introductory account of PLS we will only consider two components. Figures 4.24 and 4.25 provide the PLS score plots of the two important components. Clearly, there is a strong relationship between the X-scores (**t**) and the Y-scores (**u**), indicating that when the four factors (x´s) are changed according to the design, they exert a strong influence on the warp and strength properties of the polymers. These two score plots are similar to the fictitious score plots of Figure 4.22.

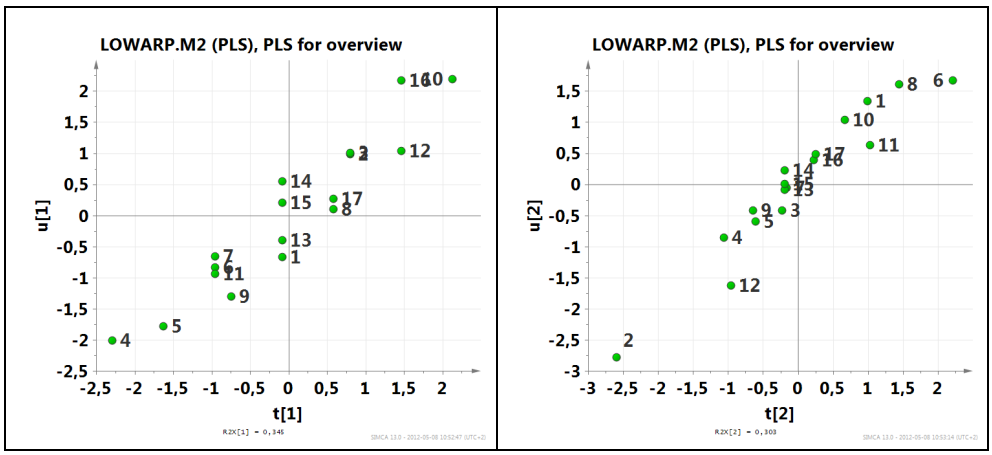


Figure 4.24: (left) PLS t_1/u_1 score plot of the model for the LOWARP dataset indicating a strong correlation between the factors and the responses. There are no outliers, i.e., off-diagonal points in the first dimension. Figure 4.25: (right) PLS t_2/u_2 score plot of the model for the LOWARP dataset. Here, we see that observations (experimental runs) 6, 11, and 12 are different, but the overall correlation is well manifested.

The objective of the data analysis was to find the combination of the factors (the four ingredients) which produces the best formulation with low warp and high strength. Such information can be obtained by considering Figures 4.26 and 4.27. The latter is a PLS weight plot ($\mathbf{w*c_2}$ is plotted against $\mathbf{w*c_1}$) showing the relationships between all four factors and 14 responses simultaneously. Patterns seen in this weight plot may then be brought back to the score plot, to see which observations are closest to fulfilling the objectives.

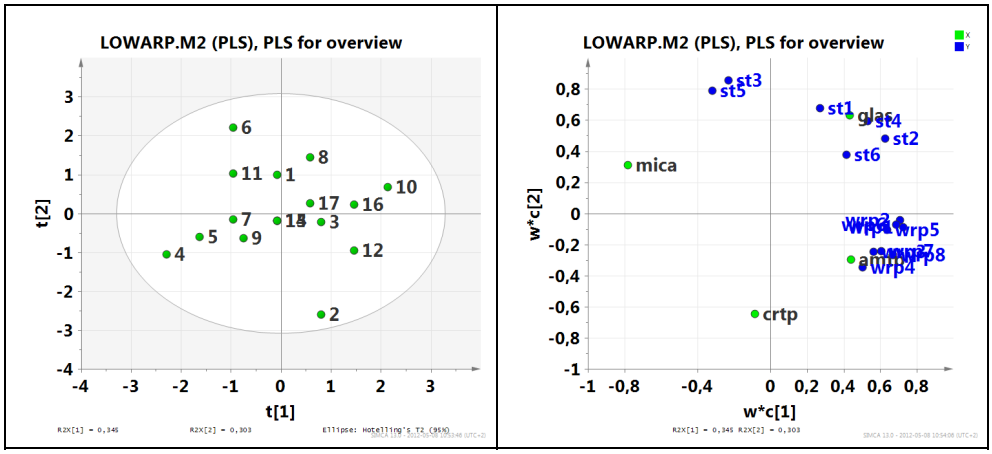


Figure 4.26: (left) PLS t_1/t_2 score plot, which is useful for understanding the relationship between the observations (the runs of the experimental design). In this case there should be a good spread among the observations, as their co-ordinates very much are regulated by the underlying design. Nevertheless this plot is relevant, if coupled to a PLS weight plot. Figure 4.27: (right) PLS weight plot of the LOWARP example. This plot reveals the inter-relatedness among the four factors (mica, amtp, crtp and glass) and the 14 responses.

Obviously, the eight warp responses are strongly correlated, as they are situated tightly together (Figure 4.27). The six strength responses, however, are less correlated, i.e., more scattered in the weight plot. In addition, the strength responses are only slightly correlated with the warp responses. We can see that glas is not a good factor for achieving the objective, as it influences both strength and warp in a similar way. Raising the amount of glas in the formulation will cause an increase in the values of all responses. In order to increase strength one should primarily focus on lowering the amount of crtp in the recipe, as this factor does not much affect warp. And to decrease warp, one should increase mica and decrease amtp.

This model interpretation reveals that observations likely to meet the demands placed on the formulation ought to be positioned in the upper left-hand part – and as far from the origin as possible – of the score plot shown in Figure 4.26. Because there are no observations in the

peripheral part of the upper left-hand area, the conclusion is that no experimental run in the design well matches the desired response profile. Consequently, the PLS model must be interrogated so as to identify experimental conditions which will provide a reasonable compromise between low warp and high strength. We can see in Figure 4.26 that such experimental runs might be observations number 6 and 11.

4.5.7 Summary of PLS projections, scores and weights

PLS modeling of the relationship between two blocks of variables can be described in different ways. Perhaps the most straightforward way is that it fits two "PCA-like" models at the same time, one for **X** and one for **Y**, and simultaneously aligns these models. The objectives are (a) to model **X** and **Y,** and (b) to predict **Y** from **X**, according to:

$$\mathbf{X} = \mathbf{1}\bar{\mathbf{x}}' + \mathbf{TP}' + \mathbf{E} \qquad \text{(eqn. 4.1)}$$

$$\mathbf{Y} = \mathbf{1}\bar{\mathbf{y}}' + \mathbf{UC}' + \mathbf{F} \quad (= \mathbf{1}\bar{\mathbf{y}}' + \mathbf{TC}' + \mathbf{G}, \text{ due to inner relation}) \qquad \text{(eqn. 4.2)}$$

In these expressions, the first terms, $\mathbf{1}\bar{\mathbf{x}}'$ and $\mathbf{1}\bar{\mathbf{y}}'$, represent the variable averages and originate from the pre-processing step. The information related to the observations are stored in the scores matrices **T** and **U**; the information related to the variables are stored in the X-loading matrix **P'** and the Y-weight matrix **C'**. The variation in the data that was left out of the modeling form the **E** and **F** residual matrices.

The difference between PCA and PLS is that the former is a maximum *variance* least squares projection of **X**, whereas the latter is a maximum *covariance* model of the relationship between **X** and **Y**. A graphical representation of the matrix relationships of PLS is shown in Figure 4.28. A detailed account of PLS is given in Appendix II.

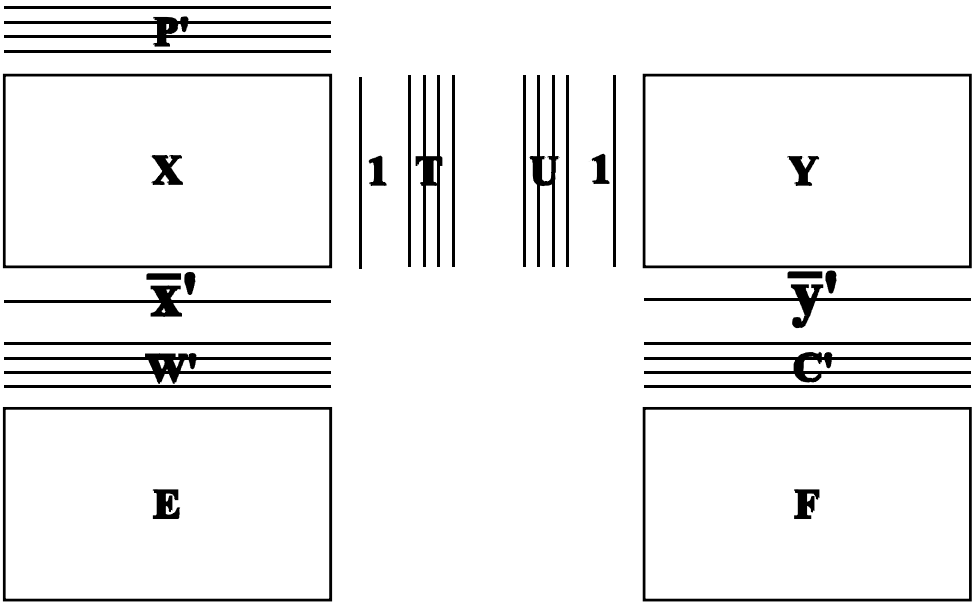


Figure 4.28: The matrix relationships in PLS. Here, $1*\bar{x}'$ and $1*\bar{y}'$ represent the variable averages and originate from the pre-processing step. The PLS scores comprise T and U, the X-loadings P', the X-weights W', and the Y-weights C'. The variation in the data that was left out of the modeling form the E and F residual matrices.

Figure 4.28 indicates the existence of the X-weight matrix **W.** This matrix contains the X-weight vectors $\mathbf{w}_a$, which show how the X-variables are linearly combined to form the score vectors $\mathbf{t}_a$. Hence, we understand which original variables dominate the new, latent variable $\mathbf{t}_a$. X-variables that are highly correlated with the Y-variables get high weights. Similarly, the Y-weights $\mathbf{c}_a$ inform us how the Y-variables are summarized by the score vector $\mathbf{u}_a$. In addition, one should observe that there are two versions of the X-weights, one denoted $\mathbf{w}_a$ and the other $\mathbf{w}_a^*$. The **w*** values relate directly to the X-matrix, whereas the **w** values refer to the residuals after the previous dimension, $\mathbf{E}_{a-1}$, instead of the X-variables themselves. The default weight plot in SIMCA shows $\mathbf{w}^*\mathbf{c}_a$.

In summary, PLS forms "new x-variables", $\mathbf{t}_a$, as linear combinations of the old ones, and thereafter uses these new **t**'s as predictors of **Y**. Only as many **t**'s (components) are formed as are predictively significant (see cross-validation in Section 4.7.3). For each component (a), the parameters, $\mathbf{t}_a$, $\mathbf{u}_a$, $\mathbf{w}_a$ (and $\mathbf{w}_a^*$), $\mathbf{p}_a$, and $\mathbf{c}_a$ are calculated by the PLS algorithm. For the *interpretation* of the PLS model, the scores, **t** and **u**, contain information about the observations and their similarities / dissimilarities with respect to the given problem and model. The weights **w*** and **c** give information about how the variables combine to form the quantitative relation between **X** and **Y**. Hence, these weights are essential for the understanding of which X-variables are important (numerically large **w***-values), which X-variables provide the same information (similar profiles of $\mathbf{w}_a^*$-values), the interpretation of the scores, **t**, etc.

4.6 PLS model interpretation

PLS provides many diagnostics which help in the model interpretation, and in the assessment of model performance and relevance. Sections 4.4 and 4.5 have concerned the PLS scores $\mathbf{t}_a$ and $\mathbf{u}_a$, what they mean and how they can be used. In the current section, we consider the variable-related parameters, notably *weights*, *coefficients*, and *VIP*. VIP is an acronym for *variable influence on projection*.

4.6.1 PLS weights

The PLS weights **w*c** give information about how the X-variables combine to form the scores **t**, the basis of the quantitative relation between **X** and **Y** (Figure 4.29). For a given PLS model, one vector of X-weights $\mathbf{w}_a^*$ and one vector of Y-weights $\mathbf{c}_a$, are obtained for each model component (a). PLS weights can be plotted in scatter, line, or column plots. Moreover, it is possible to plot the X-weights (**w***) alone, the Y-weights (**c**) alone, or both types of weights (**w*c**) in the same graph. The line plot representation is prevalent in spectroscopic applications, since it displays the "peak-like" spectral structure modeled by each component.

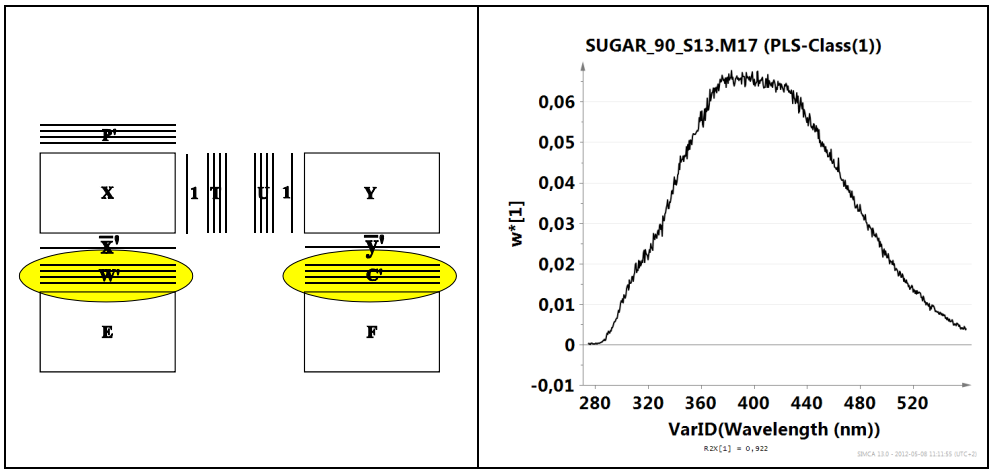


Figure 4.29: (left) To interpret a PLS model one can plot the weights, i.e. the W' (W') and C' matrices. Figure 4.30: (right) Line plot of the PLS weight w_1^* of the model for the SUGAR dataset. Here, 90 observations are modeled, and 16 unrepresentative time points are left out. The first X-weight vector closely resembles the average fluorescence spectrum (cf. Figure 4.11).*

Figures 4.30 and 4.31 show typical line plots of PLS X-weights (**w***), arising from the SUGAR dataset. (Please observe that sometimes the picture heading in these plots includes the label "loadings" although "weights" are actually being plotted. This has to do with SIMCA's internal nomenclature.) Apparently, there are two independent spectral contributions in the data. The shape of the first X-weight vector, $\mathbf{w}_1^*$, displays one strong peak, which is quite similar to the average spectrum of the raw data (cf. Figure 4.11). This

means that the spectral information that is of greatest relevance for modeling the ash content is found in the middle region of the investigated spectral range.

Moreover, the second X-weight vector ($\mathbf{w}_2^*$) also displays a peak-like structure in its left-hand part (at low wavelengths). When considering the shapes of the raw spectra, it seems as if the little "knee" displayed by some samples at low wavelengths, prior to the main peak, might be the phenomenon captured by the second PLS component.

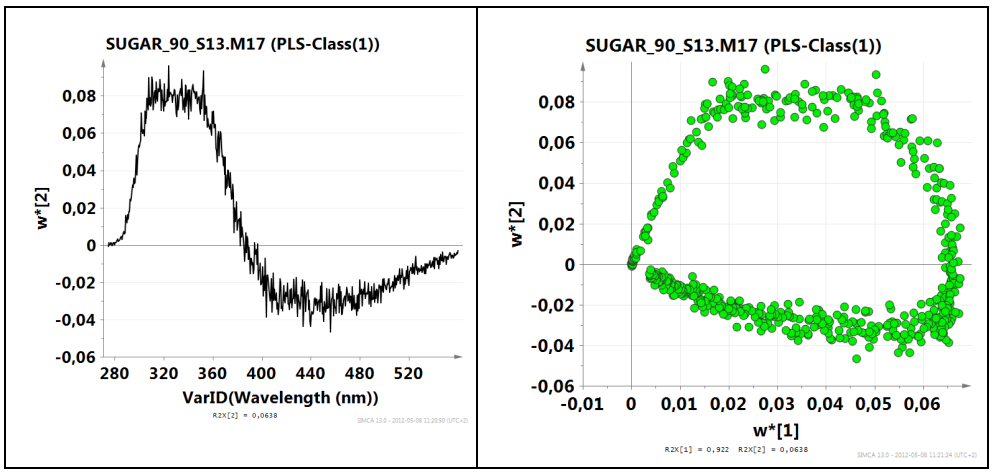


Figure 4.31: (left) Second X-weight vector of the SUGAR PLS model, w_2^*. This plot suggests a secondary spectral phenomenon. Figure 4.32: (right) Scatter plot of w_2^* versus w_1^* for the SUGAR PLS model. The curved pattern is typical for spectral data.

Scatter plots of PLS weights are not so informative for spectroscopic data. Figure 4.32 displays the scatter plot of $\mathbf{w}_2^*$ against $\mathbf{w}_1^*$ for the SUGAR dataset. We can see a strongly curved distribution of points, which is typical of spectral data. The "worm" of points originates and ends at the same place, the origin. Firstly, this is because the spectral variables (the digitized wavelengths) are strongly correlated, so the movement from one variable to another does not change the information regarding the observations. Secondly, spectral data often contain regions of low information content, which then get near-zero weights. Since such low-information regions are quite often found at the ends of the investigated spectral window, the "worm" often starts and ends at the origin.

The LOWARP dataset contains several Y-variables. Consider the PLS weight plot shown in Figure 4.33, which presents the relationships between all factors and responses at the same time. Let us interpret how the four factors combine to affect the response wrp2. First, we imagine a base line going from wrp2 through the plot origin (Figure 4.33). Second, all X-variables are projected orthogonally onto this line. The position of this projection is what counts in the model interpretation. An obvious analogy is with the behavior of a children's see-saw. When sitting close to the fulcrum (the pivotal point) you have no influence, but when sitting far out there is a substantial impact.

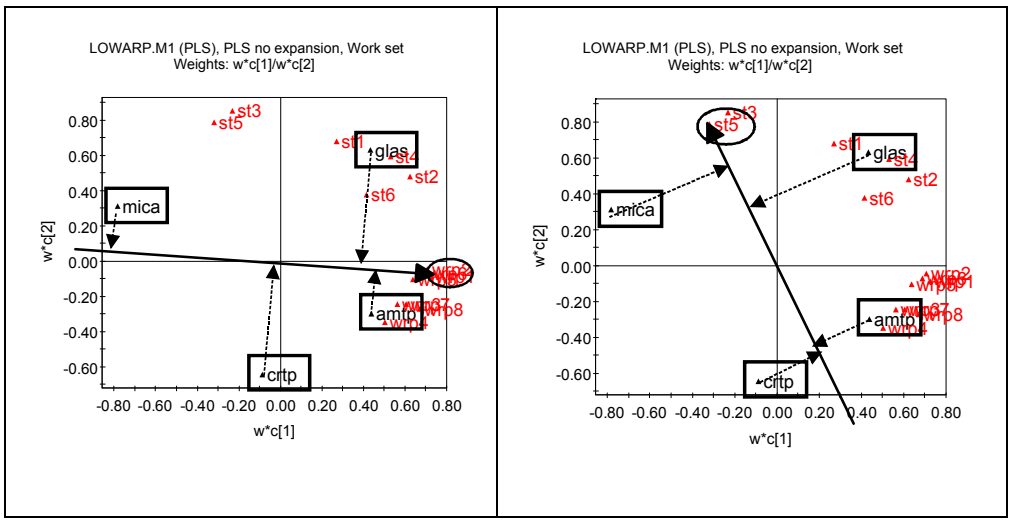


Figure 4.33: (left) Schematic illustration of how to interpret a PLS weight plot. The example pertains to the wrp2-response. Figure 4.34: (right) Same as Figure 4.33, but the drawing applies to the interpretation of the response st5.

Figure 4.33 suggests that the factors glas and amtp are positively related to wrp2, since they are projected on the same side of (0,0) as wrp2. Since the objective is to decrease wrp2, we realize that we should lower the amounts of amtp and glas in the formulation. These two factors are important for wrp2, but even more influential is mica. This is understood from the fact that mica is projected furthest away from the origin. When mica increases wrp2 is predicted to decrease. Finally, the factor crtp has almost no impact on wrp2, because it is projected close to (0,0).

The situation changes drastically when we examine a response that is only slightly correlated with wrp2, for instance st5 (Figure 4.34). Now, crtp becomes an important factor that should be decreased to increase this strength response. The same holds true for amtp. Mica is more influential than glas, but they should both be raised in order to enhance st5.

4.6.2 PLS regression coefficients

A PLS solution given in the latent variable framework (with scores, weights, etc.) may be re-expressed as a regression model consisting of PLS regression coefficients, $\mathbf{B}_{PLS}$, according to

$$\mathbf{Y} = \mathbf{1}\bar{\mathbf{y}}' + \mathbf{X}\mathbf{B}_{PLS} + \mathbf{F} \qquad \text{(eqn. 4.3)}$$

The relationship between the PLS regression coefficients and the PLS weights is given by:

$$\mathbf{B}_{PLS} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} = \mathbf{W}^{*}\mathbf{C} \qquad \text{(eqn. 4.4)}$$

PLS coefficients are of interest because they simplify the model interpretation when there are several components (> 4-5) in the model. Their advantage is that the analyst obtains only one vector of concise model information per response, rather than several vectors of weights. The disadvantage of the coefficients is that information regarding the correlation structure among the responses is lost. This information is preserved by the PLS weights.

In order to illustrate PLS regression coefficients, we shall look at the LOWARP dataset. We will not show any results related to the SUGAR example, as this dataset contains only one response, and there would be little to gain by switching from weights to coefficients in the interpretation of the calibration model. Since the LOWARP dataset contains 14 responses, the two-dimensional PLS model obtained can be expressed as 14 sets of regression coefficients. For practical reasons, however, we only visualize regression coefficients for 4 responses (Figures 4.35 – 4.38), i.e., those for wrp1, wrp2, wrp4, and st4.

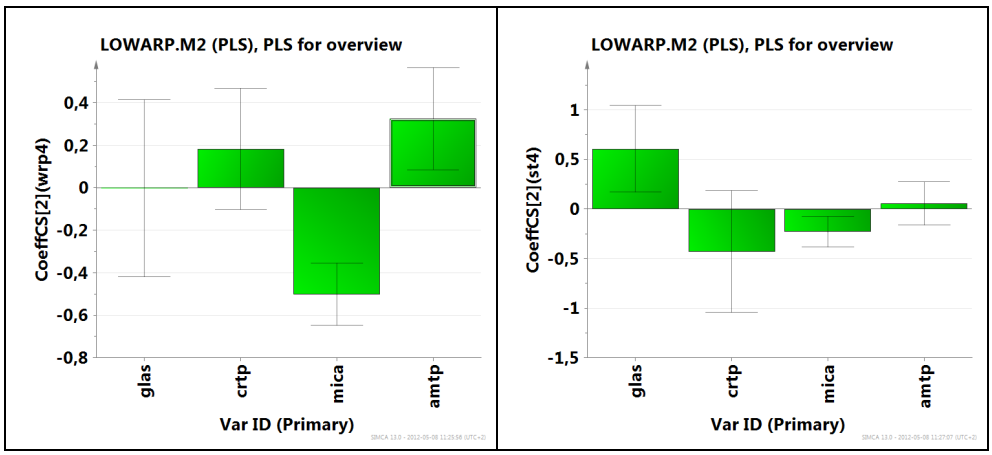


Figure 4.35: (left) Regression coefficients of scaled and centered variables for the response wrp4.
Figure 4.36: (right) Regression coefficients of scaled and centered variables for the response st4.

Figures 4.35 and 4.36 display regression coefficients for two uncorrelated responses, wrp4 and st4. That these responses are indeed uncorrelated can be seen in the PLS weight plot of Figure 4.27. Hence, it is natural that the two responses exhibit markedly different regression coefficient profiles. In the coefficient plots, the sizes and signs of the coefficients relating to centered and scaled variables indicate the influence of each model term. And, the statistical significance of each coefficient is indicated with 95% confidence intervals. Thus, wrp4 is most sensitive to changes in mica and amtp, whereas st4 is predominantly affected by glas and crtp, and to some extent mica.

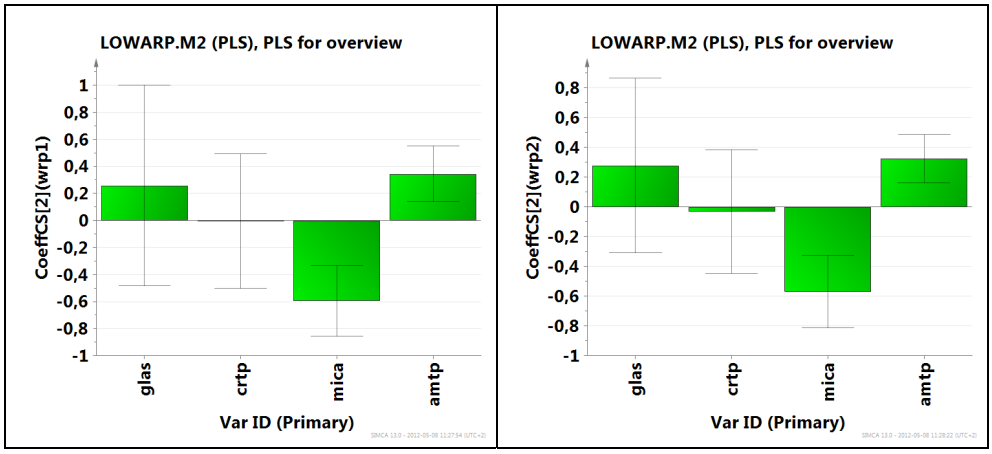


Figure 4.37: (left) Regression coefficients of scaled and centered variables for the response wrp1.
Figure 4.38: (right) Regression coefficients of scaled and centered variables for the response wrp2.

Next, we consider Figures 4.37 and 4.38, where similar profiles of coefficients are encountered. The close resemblance of these two plots indicates that wrp1 and wrp2 are strongly correlated, which is also corroborated by the fact that these two responses are situated almost on top of one another in Figure 4.27. Obviously, mica is the most important factor, but the influences of glas and amtp cannot be neglected. To diminish polymer warp we should primarily focus on increasing mica in the mixture.

In summary, PLS regression coefficients offer a compact representation of the X-Y relations of a PLS model, which facilitates model overview and interpretation. There are several ways to express the coefficients, but scaled and centered coefficients are most often used. Alternative representations are discussed in Appendix II. The paper by Burnham and co-authors provides further insight into the interpretation of regression coefficients under a latent variable regression model [Burnham, *et al.*, 2001].

Finally, in spectroscopic applications it is sometimes desirable to rotate the PLS regression coefficients such that they reflect the pure spectral profiles of the analytes in the samples. This is discussed in Section 5.4.4.

4.6.3 The variable influence on projection, VIP, parameter

Interpreting a PLS model with many components and a multitude of responses can be a complex task. A parameter which summarizes the importance of the X-variables, both for the X- and Y-models, developed by Wold in 1993 [Wold, *et al.*, 1993a], is called the *v*ariable *i*nfluence on *p*rojection, VIP. The details of VIP are given in Appendix II. For the moment, it suffices to state that VIP is a weighted sum of squares of the PLS weights, **w***, taking into account the amount of explained Y-variance in each dimension. Its attraction lies in its intrinsic parsimony; for a given model and problem there will always be only *one* VIP-vector, summarizing all components and Y-variables (Figure 4.39).

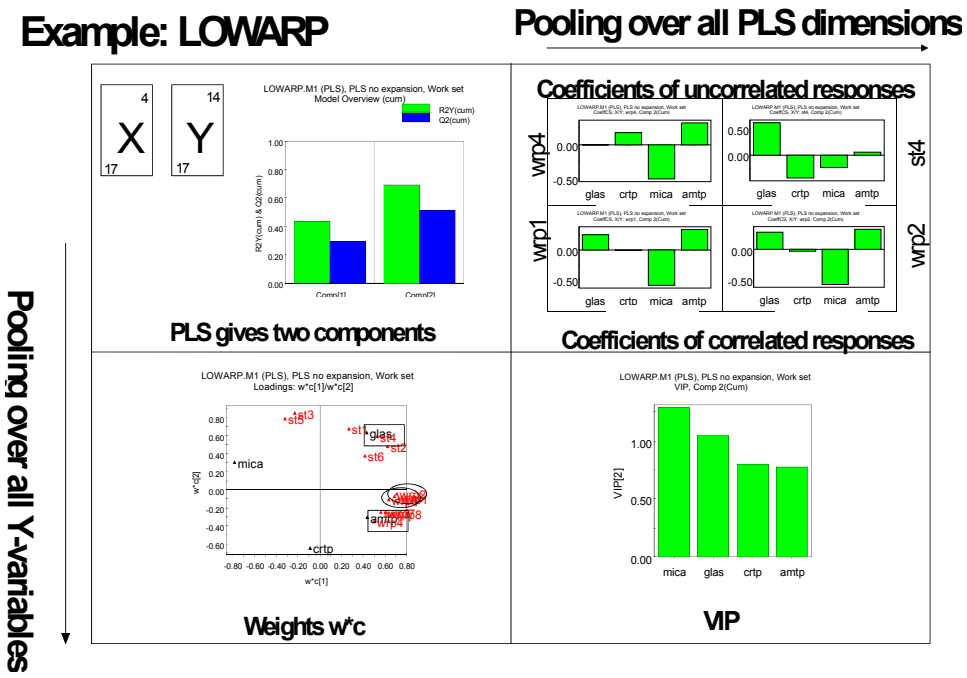


Figure 4.39: The weights, coefficients, and VIP parameters of the LOWARP example. PLS analysis of the LOWARP data yielded a two-dimensional model, which may be interpreted by means of the weights (lower left). Alternatively, the PLS model may be interpreted by means of the coefficients. With 14 responses, there are 14 sets of coefficients, although we only display four (upper right). This kind of model information is obtained by pooling over all PLS components. An even more condensed basis for model interpretation is given by VIP, which involves pooling over all components and all responses.

Figure 4.39 illustrates the relation between weights, coefficients, and VIP, for the LOWARP example. The PLS analysis of the LOWARP data yielded a two-dimensional model, which may be interpreted by means of the weights (lower left-hand part of Figure 4.39). This type of model information is obtained by pooling over all 14 responses. For illustrative purposes, the locations of two uncorrelated responses (wrp4 and st4) are boxed, and the positions of two correlated responses (wrp1 and wrp2) are indicated by two small ellipses.

Alternatively, the PLS model may be interpreted by means of the coefficients. With 14 responses, it is necessary to survey 14 sets of coefficients, although only four such sets are displayed. These four sets of coefficients relate to the four responses wrp1, wrp2, wrp4, and st4. As expected, the two correlated responses (wrp1 and wrp2) demonstrate very similar coefficient profiles, whereas the corresponding profiles for the two uncorrelated responses (wrp4 and st4) differ.

Furthermore, VIP compactly summarizes the importance of the X-variables, both for the X- and Y-model parts. For a given problem and model, there is only one set of VIP-values.

Observe that VIP is a squared function of the PLS weights, and therefore only positive values will be encountered. Predictors with a large VIP, larger than 1, are most influential for the model. This means that mica and glas contribute most strongly to the modeling of the 14 warp and strength responses. In routine PLS analysis, it is recommended to couple a study of VIP to an inspection of either weights or coefficients.

VIP can be a fruitful alternative for model interpretation, especially when the PLS model has many dimensions. Figure 4.40 shows the VIP-plot for the SUGAR dataset. For comparison, we re-display in Figure 4.41 the corresponding PLS weight plot for the first model dimension.

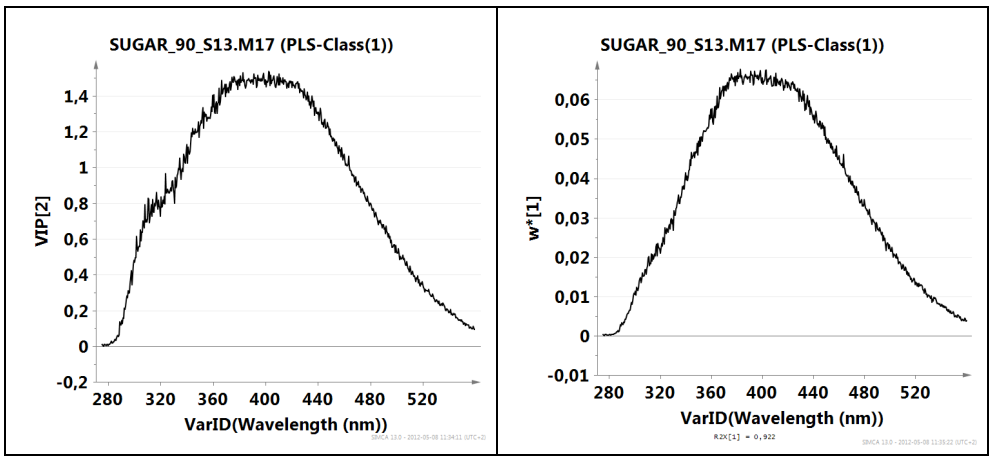


Figure 4.40: (left) VIP plot for the PLS model of the SUGAR dataset. Here, 90 observations are modeled and 16 unrepresentative time points are left out. Figure 4.41: (right) Line plot of the PLS weight w_1^* of the first component of the model for the SUGAR dataset. The first weight vector closely resembles the average fluorescence spectrum (cf. Figure 4.11).

Interestingly, the VIP-values may be used for a mild variable selection. In e.g. QSAR when dealing with many X-variables, a column plot of VIP often looks like a stair case (Figure 4.42). Thus, a plot of VIP might point to one or several thresholds, which might be tested for their ability to discriminate between important and unimportant predictors. In most cases, a cut-off for VIP around 0.7 - 0.8 works well. More thorough discussions regarding variable selection and model pruning are given in Chapters 7 – 11, and 18.

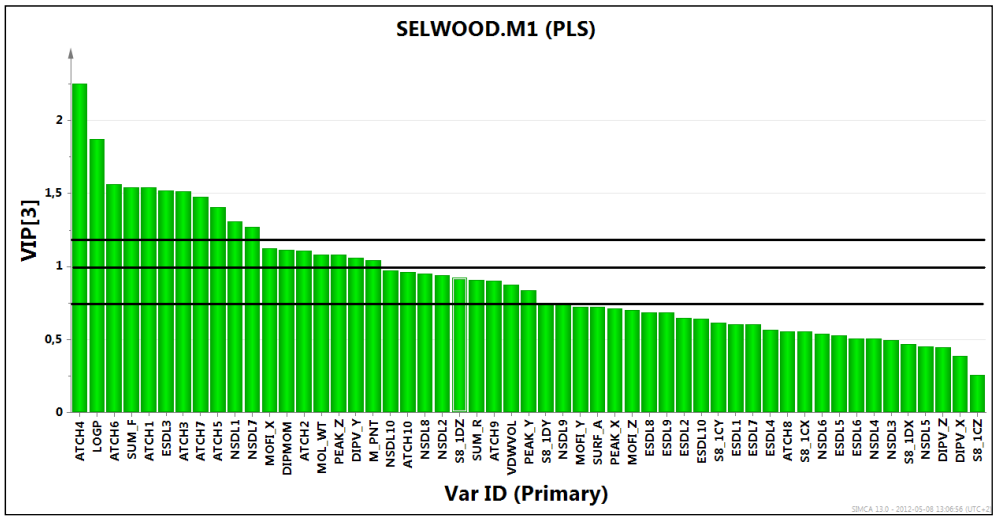


Figure 4.42: VIP plot of a three-dimensional PLS model, indicating different thresholds which might be tested in variable selection. Note, however, that variable selection should be carried out with caution, since there are strong correlations among the X-variables.

4.7 Additional PLS diagnostics

PLS offers a number of useful model parameters and diagnostic tools. Many of these tools were introduced in Chapter 3 in connection with PCA. We will now briefly review these tools in relation to PLS. The current section is organized as follows: first we discuss observation diagnostics (Section 4.7.1), then variable diagnostics (Section 4.7.2), and finally model diagnostics (Section 4.7.3).

4.7.1 Observation diagnostics – Are there non-linearities and/or outliers in the data?

Outliers may be either strong or moderate. In PLS (and PCA), the former are found by inspecting the scores, and the latter by looking at the residuals. First, we consider the diagnostic possibilities offered by the PLS scores. An observation may be an outlier in **X**, in **Y**, and/or in the relationship between **X** and **Y**. The latter case is schematically illustrated in Figures 4.43 and 4.44.

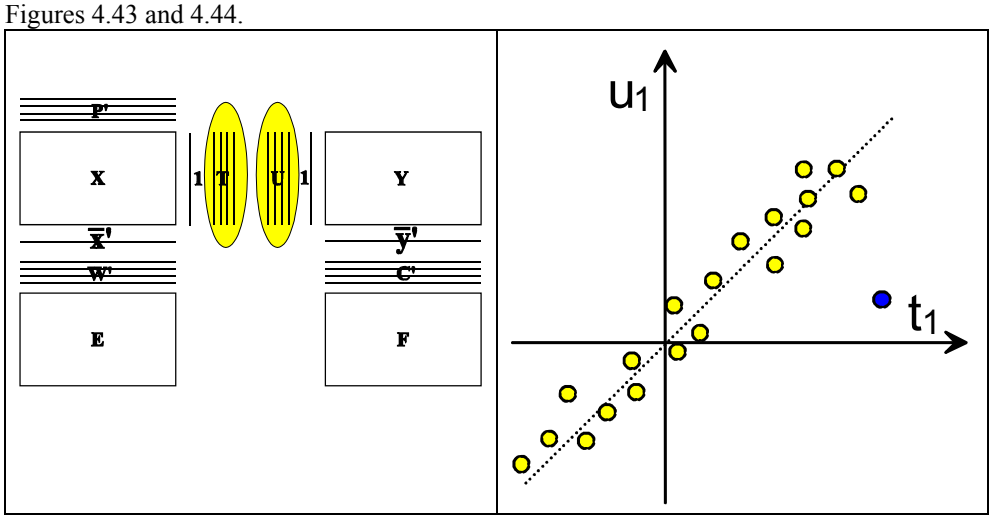


Figure 4.43: (left) For uncovering outliers among observations, it is informative to make scatter plots of the PLS scores, t and u. Figure 4.44: (right) An observation may be an outlier in the X/Y relationship.

In the LOWARP case, the PLS model contained two components. Hence, four score plots are conceivable, namely $\mathbf{t_1/u_1}$, $\mathbf{t_2/u_2}$, $\mathbf{t_1/t_2}$, and $\mathbf{u_1/u_2}$. PLS score plots of the t/u-type are used to uncover deviations from the dominating **X/Y** correlation structure (Figures 4.45 and 4.46).

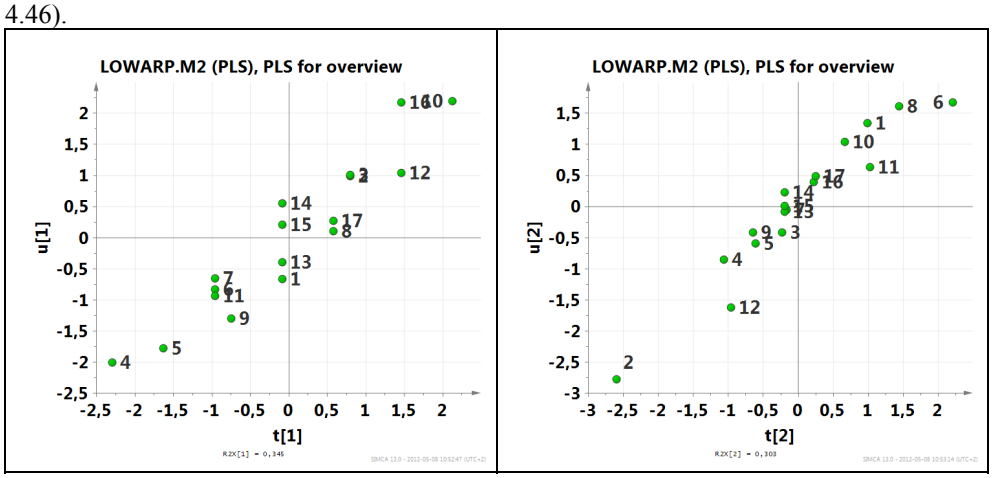


Figure 4.45: (left) PLS t_1/u_1 score plot of the model for the LOWARP dataset indicating a strong correlation between the factors and the responses. There are no outliers, i.e., off-diagonal points in the first dimension. Figure 4.46: (right) PLS t_2/u_2 score plot of the model for the LOWARP dataset. Here, we see that observation (experimental run) numbers 6, 11, and 12 are somewhat different from the others, but the overall correlation is strong.

As well as detecting outliers in the **X**/**Y** relationship, the PLS **t**/**u** scores are useful in identifying departures from linearity between **X** and **Y**. In this context, curvature between **X** and **Y** may manifest itself in two ways. Figure 4.47 depicts the first type of non-linearity, indicating the need for transformation of response data. A real example of what the $\mathbf{t}_1/\mathbf{u}_1$ score plot may look like in the untransformed situation is given in Figure 4.48, and the result after log-transformation of the response variable is given in Figure 4.49. Clearly, the transformation of the response is warranted.

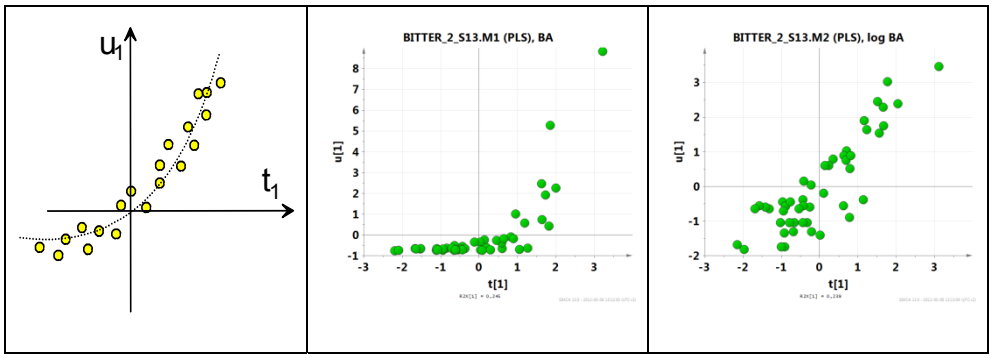


Figure 4.47: (left) Non-linearity between X and Y is often seen in the first PLS component. This phenomenon can be visualized by plotting t_1 and u_1 against each other. Figure 4.48: (middle) QSAR example where a non-linear relationship is found between t_1 and u_1. Figure 4.49: (right) By log-transforming the response variable (y), the inner relation in the PLS model becomes more nearly linear.

The other main type of non-linearity between **X** and **Y**, expressed by, say, $\mathbf{t}_1$ and $\mathbf{u}_1$, involves a curvature with inverse arching (Figure 4.50). In this case a possible remedy might be a transformation of **X**. Figure 4.51 shows what a $\mathbf{t}_1/\mathbf{u}_1$ plot may look like in a process dataset from a continuous process. As shown by Figure 4.52, the curved behavior disappears after log-transforming the predictor variables and refitting the PLS model.

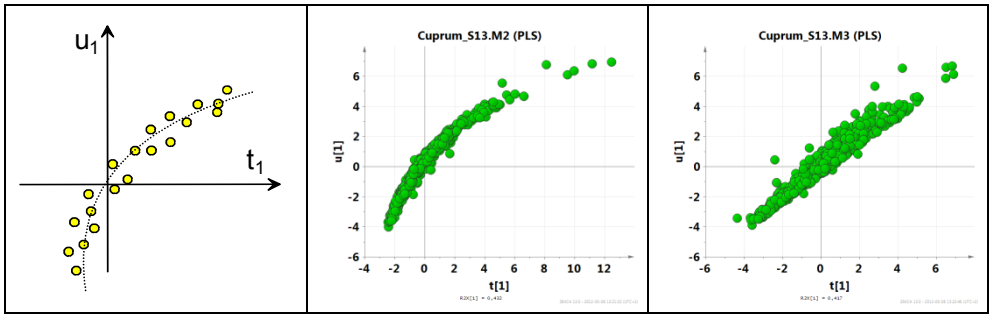


Figure 4.50: (left) The curvature between X and Y indicates the need for transformation of the X-variables. Figure 4.51: (middle) Process dataset where the inner relation is non-linear. Figure 4.52: (right) By log-transforming the predictor variables (X), the inner relation in the PLS model becomes linear.

Moreover, by plotting the t-vectors against each other and the u-vectors against each other, we may examine the patterns in the X- and Y-spaces separately. Figure 4.53 shows $\mathbf{t}_1$ and $\mathbf{t}_2$ of the model for the LOWARP dataset, and Figure 4.54 displays the corresponding $\mathbf{u}_1/\mathbf{u}_2$ score plot. Here, we find no apparent outliers either in **X** or **Y**.

In more general terms, a t/t score plot is useful for finding deviations in the X-data and a u/u plot for finding deviations in the Y-data. Deviating observations in **X** often correspond to changing operating conditions in a process industry, modified composition of chemical samples in analytical chemistry, or atypical properties of chemical compounds in QSAR. Similarly, observations that deviate in **Y** often represent fluctuating or degrading quality in manufacturing processes, non-representative samples for multivariate calibration, or compounds deviating from the common mode of biological action in a QSAR dataset.

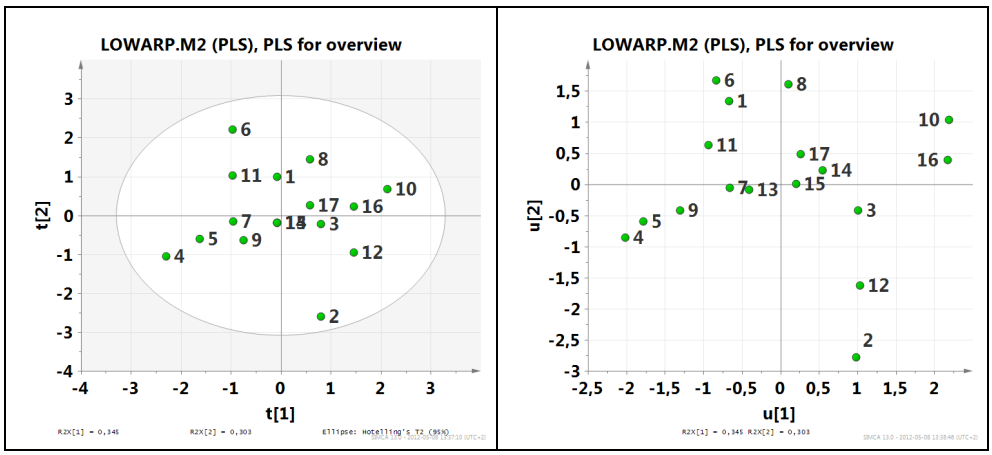


Figure 4.53: (left) PLS t_1/t_2 score plot of the model for the LOWARP dataset. All observations are well spread in the score space, which is to be expected since the dataset results from a mixture design. Figure 4.54: (right) PLS u_1/u_2 score plot of the model for the LOWARP dataset. There are no obvious groupings among the observations in the Y-data. Rather, there is good agreement between the positions in the t_1/t_2 plot and in the u_1/u_2 plot. The u/u plot will reveal strong deviations in the Y-data, for instance quality fluctuations in process monitoring, or differing biological performance of compounds in QSAR.

Now, we shall focus on the PLS residuals, i.e., matrices **E** and **F** (cf. Equations 4.1 and 4.2). As discussed in Chapter 3, an inspection of the residuals enables moderate outliers to be detected. This is graphically illustrated in Figure 4.55, where "beer cans" have been inserted to surround the data swarms in X- and Y-space. These "beer cans" correspond to tolerance volumes, which indicate the typical range of variation among the observations, and allow formal testing of whether any new observation is similar to those used for building the model. The **E** and **F** residual matrices are thus related to the sizes of these "beer cans".

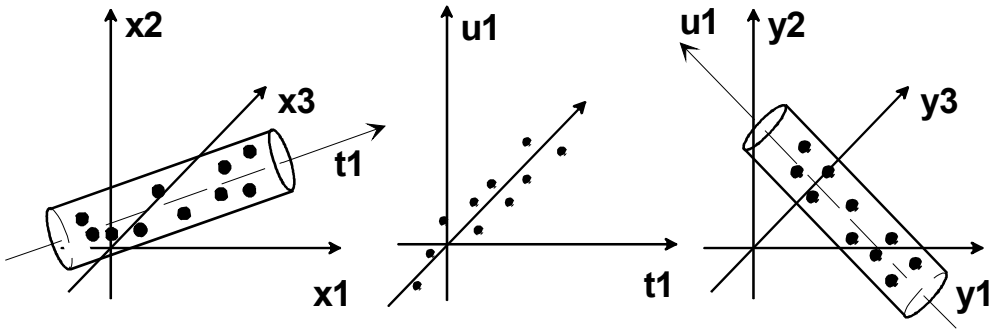


Figure 4.55: In PLS, the unexplained variation, the residuals, can be used to construct tolerance volumes around the data swarms. Such tolerance volumes indicate the typical range of variation among the observations, and allow formal testing of whether a new observation is similar to those used for building the model. The residual elements in the matrices E and F are related to the diameters of the "beer cans". Note that the geometrical interpretation of the tolerance volumes as "beer cans" is only appropriate when considering a one-dimensional PLS model.

Moderate outliers are seen in plots of DModX/DModY, which show each observation's *distance to the model in X/Y-space*. DModX and DModY are based on a row-wise summation of the elements of the residual matrices **E** and **F** (Figure 4.56) and are equivalent to the row residual standard deviation. For a one-dimensional PLS model this leads to the formation of the "beer-can" like tolerance volumes in **X** and **Y** (Figure 4.57).

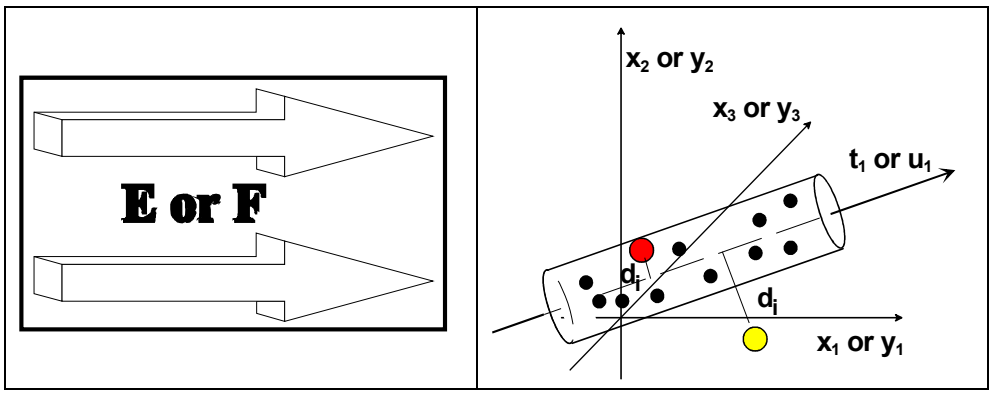


Figure 4.56: (left) In the computation of DModX and DModY, the elements of E and F are summarized row-wise. Figure 4.57: (right) A tolerance volume enclosing a point-swarm can be used as a diagnostic tool to evaluate whether new observations are similar or dissimilar to the training set members. In the plot, the yellow observation would be considered as deviating from the model and hence likely be a moderate outlier.

Plots of DModX and DModY, with and without moderate outliers present, are shown in Figures 4.58 – 4.61. The first pair of plots, Figures 4.58 and 4.59, relate to the LOWARP example, where no moderate outliers can be detected, neither in **X** nor in **Y**. The size of the tolerance volume in X is indicated by the solid line, D_{crit}, in Figure 4.58. All 17 observations are located inside the tolerance volume. The default critical distance in SIMCA corresponds to the 0.05 probability level, but this may be changed according to the preferences of the user. Also note that no D_{crit} is given for DModY.

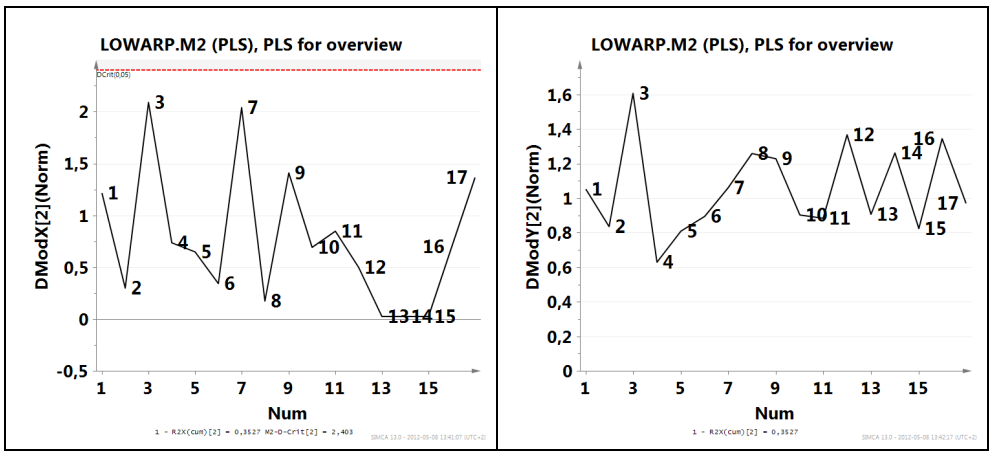


Figure 4.58: (left) DModX plot for the PLS model of the LOWARP dataset. No moderate outliers are present, as all observations are situated within the critical distance. Figure 4.59: (right) DModY plot for the PLS model of the LOWARP dataset. No moderately deviating observations are detected.

Now consider the pair of DModX/Y plots given in Figures 4.60 and 4.61. These two plots concern data from two processes. There are between 5 and 10 moderate outliers located outside the critical distance at the 0.05 probability level for the first dataset (Figure 4.60). Likewise, the response data of the second dataset suggest the existence of several moderate outliers (Figure 4.61).

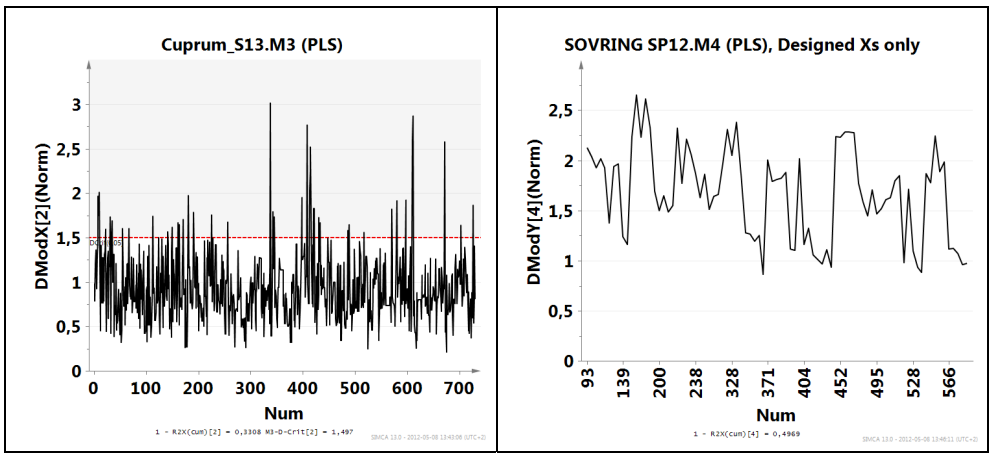


Figure 4.60: (left) DModX plot of a process with moderate outliers outside the critical distance. Figure 4.61: (right) DModY plot of a process with moderate outliers. Several observations have response profiles which differ from the majority of observations, as indicated by comparatively large DModY values.

The calculation of DModX can be formalized as follows. The residual observation variance, *S2OX*, (for observation *i*) is

$$S2OX = \sum_k e_{ik}^2 / DF \qquad \text{(eqn. 4.5)},$$

where DF represents the number of degrees of freedom. The residual observation variance can be converted to the absolute distance DModX as

$$DModX_{abs} = (S2OX)^{1/2} \qquad \text{(eqn. 4.6)},$$

or the normalized distance DModX as

$$DModX_{norm} = [S2OX/variance\ (E)]^{1/2} \qquad \text{(eqn. 4.7)}.$$

In order to arrive at the corresponding formulas for DModY, the residual e_{ik} is replaced by the residual f_{im}, according to:

$$S2OY = \sum_m f_{im}^2 / DF \qquad \text{(eqn. 4.8)},$$

$$DModY_{abs} = (S2OY)^{1/2} \qquad \text{(eqn. 4.9)},$$

$$DModY_{norm} = [S2OY/variance\ (F)]^{1/2} \qquad \text{(eqn. 4.10)}.$$

A more thorough account of these parameters is provided in Appendix II.

4.7.2 Variable diagnostics – Which predictors are important and which responses are well explained?

In Section 3.5.2 it was explained that the elements of a residual matrix may be summarized column-by-column. If this is performed for the X-residuals in matrix **E** (Figure 4.62), it is possible to compute the *explained variation* (*R2VX*) of a variable. (This quantity is often denoted as just R^2, but here we use the SIMCA notation R2VX). *R2VX* ranges from 0 (no explanation) to 1 (complete explanation), and reveals how predictors are explained by the model. Figures 4.63 and 4.64 demonstrate how the explained variation of each predictor in the LOWARP dataset increases when going from one to two components in the PLS model.

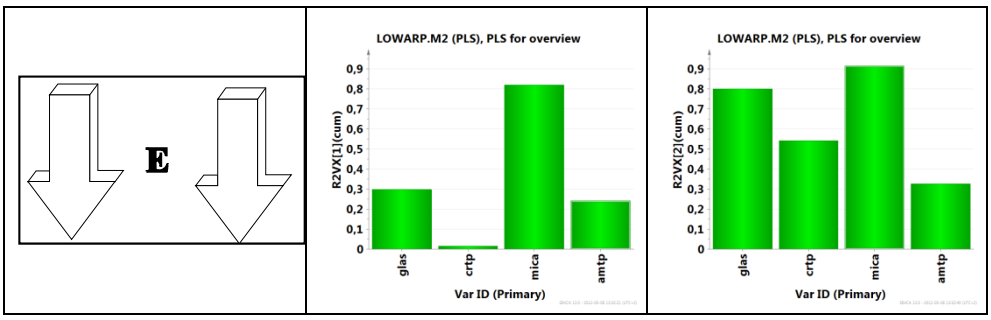


Figure 4.62: (left) In the computation of the explained variation of predictors, R2VX, the residuals of matrix E are summarized column-wise. Figure 4.63: (middle) Explained variation of the four predictors in the LOWARP dataset after one PLS component. Mica is the most important factor. Figure 4.64: (right) Same as in Figure 4.63, but results are shown for two PLS components. The two factors glas and crtp are now much better described by the model.

Moreover, the *explained variance* ($R2VX_{adj}$) is simply the explained variation divided by the number of degrees of freedom (DF).

Initially, the residual variable variation, *SSVX*, is computed as

$$SSVX = \Sigma_i e_{ik}^2 \qquad \text{(eqn. 4.11)},$$

which becomes the residual variable variance, *S2VX*, by dividing by the degrees of freedom

$$S2VX = \Sigma_i e_{ik}^2/DF \qquad \text{(eqn. 4.12)}.$$

The cumulative explained variation and explained variance after the A´th component of a variable are given respectively by

$$R2VX\ (cum) = 1 - SSVX[A]/SSVX[0] \qquad \text{(eqn. 4.13)}$$

$$R2Vxadj\ (cum) = 1 - S2VX[A]/S2VX[0] \qquad \text{(eqn. 4.14)},$$

where *A* represents the number of PLS components. A more thorough description of these parameters is found in Appendix II. Also note that it is possible to calculate *R2X*- and *R2Xadj*-values pertaining to the complete X-matrix (and not only to the individual predictor variables).

In addition to the X-residuals in matrix **E**, PLS also provides Y-residuals stored in matrix **F**. These may also be summarized column-wise (Figure 4.65). The sizes of the Y-residuals show which responses are well accounted for by the PLS model. Quantitatively, this information is given by the explained variation, *R2VY*. Figures 4.66 and 4.67 display how *R2VY* changes for the 14 responses of the LOWARP dataset when using respectively one and two components in the PLS model. Generally, after one component the warp responses are better modeled than the strength responses. The second component improves the description of both types of response. In particular, one may note the strong improvement seen for st3.

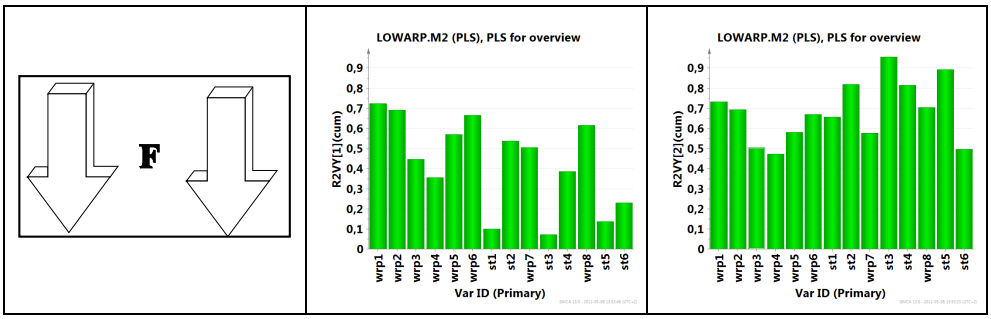


Figure 4.65: (left) In the computation of the explained variation of responses, R2VY, the residuals of matrix F are summarized column-wise. Figure 4.66: (middle) Explained variation of the 14 responses in the LOWARP dataset after one PLS component. Many warp-related responses are well explained after the first PLS component. Figure 4.67: (right) Same as in Figure 4.66, but for two PLS components. Almost all responses are now modeled well. Only three responses (wrp3, wrp4, and st6) exhibit R2VY-values around 0.5.

The computation of *R2VY* is shown below. First, the residual response variable variation, *SSVY*, is computed as

$$SSVY = \Sigma_i f_{im}^2 \quad \text{(eqn. 4.15)},$$

which, by dividing by the degrees of freedom, gives the residual response variable variance, *S2VY*,

$$S2VY = \Sigma_i e_{ik}^2/DF \quad \text{(eqn. 4.16)}.$$

The cumulative explained variation and explained variance after the A´th component of a response variable are given by

$$R2VY\ (cum) = 1 - SSVY[A]/SSVY[0] \quad \text{(eqn. 4.17)}$$

$$R2VYadj(cum) = 1 - S2VY[A]/S2VY[0] \quad \text{(eqn. 4.18)}.$$

Here, *A* denotes the number of PLS components. A more thorough description of these parameters is found in Appendix II. Moreover, it is possible to calculate *R2Y*- and *R2Yadj*-values for the complete Y-matrix (and not just for the individual response variables).

4.7.3 Model diagnostics – How many PLS components should be used?

In order to determine the appropriate number of components in a PLS model, the technique of cross-validation is useful. By "appropriate" we mean that we require a model with an optimal balance between fit and predictive ability (Figure 4.68). We recall from Chapter 3 that the fit tells how well we are able to mathematically reproduce the data of the training set, and the predictive ability how reliably we can predict the outcome of tomorrow's experiments. A quantitative measure of the *goodness of fit* is given by the parameter R^2 (= the explained variation). The predictive ability, on the other hand, is given by the *goodness of prediction* parameter Q^2 (= the predicted variation). *In the context of PLS, the terms R^2 and Q^2 generally refer to the model performance of the Y-data, the responses, rather than of the X-data, the predictors, as is the case in PCA.*

Usually, the R^2Y- and Q^2Y-parameters vary differently with increasing model complexity (Figure 4.68). R^2Y is inflationary and rapidly approaches unity as model complexity (number of model parameters) increases. Hence, it is not sufficient to have a high R^2Y. The goodness of prediction, Q^2Y, on the other hand, is not inflationary and will not automatically approach 1 with increasing model complexity. Commencing with a very simple model, Q^2Y will increase with increasing model complexity (Figure 4.68). At a certain degree of complexity, however, the predictive ability will not improve any further. The value of Q^2Y reaches a plateau, and will subsequently head downwards (indicating that the predictive ability is degrading). Thus, there is a trade-off between fit and predictive ability. And it is this zone, where we have a balance between predictive power and reasonable fit, that we wish to identify.

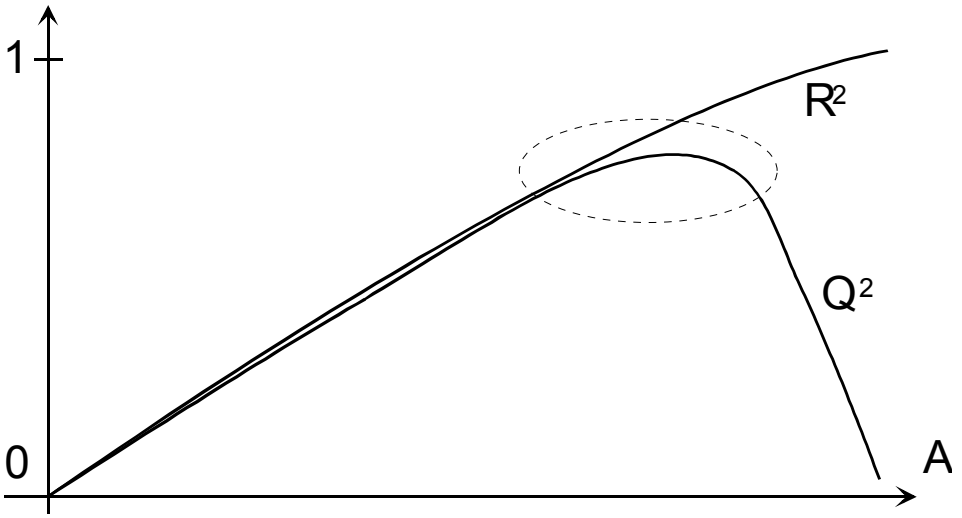


Figure 4.68: The trade-off between the goodness of fit, R^2Y, and the goodness of prediction, Q^2Y. The vertical axis corresponds to the amount of explained or predicted variation, and the horizontal axis depicts the model complexity (number of terms, number of latent variables, etc). At a certain model complexity, we obtain the most valid model, exhibiting the optimal balance between fit and predictive ability.

The R^2Y- and Q^2Y-values are informative regarding model performance and evolution. Concerning the LOWARP application, Figure 4.69 displays the evolution of R^2Y and Q^2Y with increasing model complexity. In this case, the predictive ability (Q^2Y) does not increase appreciably beyond the second PLS component.

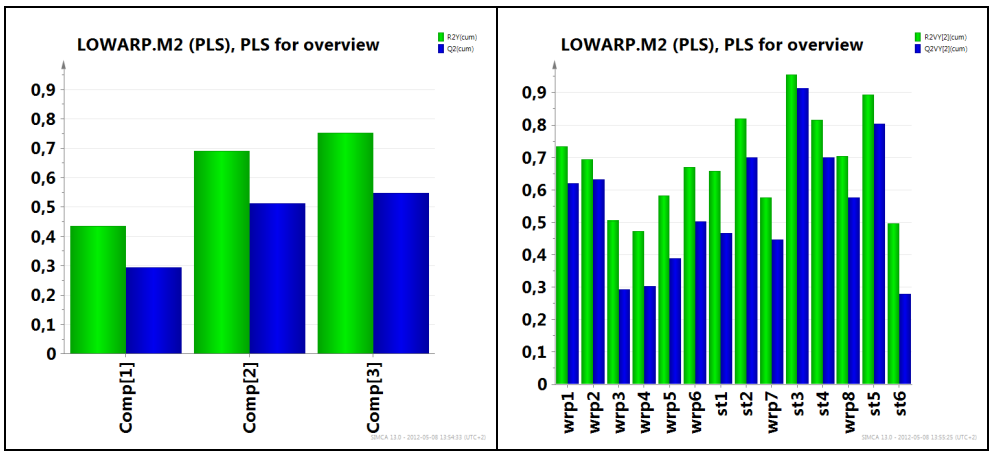


Figure 4.69: (left) Overview of the PLS model for the LOWARP dataset. The plot shows how the overall R^2Y- and Q^2Y-statistics change as a function of increasing model complexity. Here, two components appear appropriate, as Q^2Y increases only marginally with the introduction of the third component. Figure 4.70: (right) The overall R^2Y and Q^2Y-values may be decomposed into R^2's and Q^2's relating to the individual responses. The situation after two model components is depicted. Most variables are rather well modeled after two PLS components, but there are some exceptions. Particularly, one should analyze the variables wrp4 and st6 in more detail to understand why they are so poorly described, and the relatively large gaps between their R^2Y and Q^2Y.

By investigating the R^2Y*'s* and Q^2Y*'s* of the individual responses it is possible to see which responses are well modeled, and which are not (Figure 4.70). Figure 4.70 presents the situation after two PLS components. Apparently, the situation is good or very good for responses like st3, st4, and st5, which all display Q^2Y*'s* above 0.7. However, there are also less well modeled responses, such as wrp3, wrp4, and st6, which have Q^2Y*'s* below 0.4. Nevertheless, the overall picture shows a strong, quantitative relationship between the factors and the responses.

As mentioned at the beginning of this section, we use cross-validation [Wold, 1978] not only to estimate the overall Q^2Y of the PLS model for the LOWARP dataset, but also to

compute the Q^2Y's of the individual response variables. Basically, CV is performed by dividing the data into a number of groups (SIMCA default = 7) and then developing a number of parallel models from reduced data with one of the groups deleted (Figure 4.71). It should be noted that increasing the number of CV groups to N, i.e., the so called leave-one-out approach, is not recommended, because the estimated Q^2Y then becomes too similar to R^2Y.

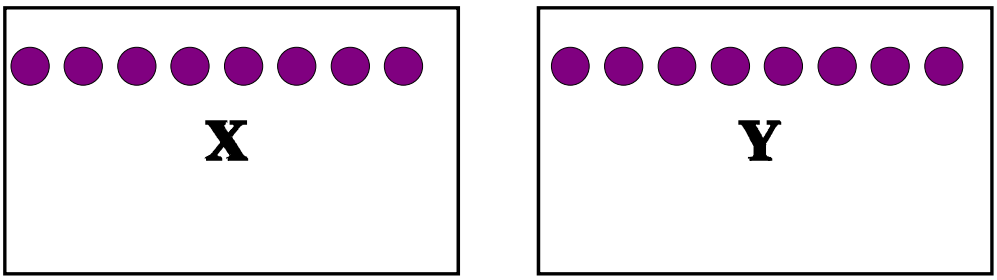


Figure 4.71: In PLS data are divided into G groups, typically between five and ten groups, and temporarily deleted in a row-wise fashion. A model is established for the data devoid of one group. The Y-data of each deleted row are predicted by the model and the predictive residual sum of squares (PRESS) is calculated. This procedure is repeated G times, followed by the summation of all partial PRESS-values to give an overall PRESS-value. If a new PLS component, PLS_a, enhances the predictive power compared with the previous component, PLS_{a-1}, the new component, PLS_a, is retained in the model.

After developing the reduced model, the omitted data are used as a test set, and the differences between actual and predicted Y-values are calculated for these data points. The sum of squares of these differences from all the parallel models are used to form *PRESS* (predictive residual sum of squares). This is a measure of the predictive ability of the model:

$$PRESS = \Sigma (y_{im} - \hat{y}_{im})^2 \qquad \text{(eqn. 4.19).}$$

When CV is used in the sequential mode, $PRESS_a/SS_{a-1}$ is evaluated after each component, and a component is judged significant if this ratio is smaller than around 0.9 for at least one of the y-variables (sharper bounds can be obtained from the results of Wakeling and Morris [Wakeling and Morris, 1993]). Here, SS_{a-1} denotes the (fitted) residual sum of squares before the current component (index a). The calculations continue until a component is non-significant. Alternatively, however, one can calculate *PRESS* for each component up to, say 10 or 15 components, and use the model which gives the lowest *PRESS/(N-A-1).* This "total" approach is computationally much more taxing, and the practical difference from the "sequential" CV results is small.

Both with the “sequential” mode and the "total" mode, a PRESS is calculated for the final model with the estimated number of significant components. This is often re-expressed as Q^2Y (the "cross-validated R^2Y"), a statistic which is similar to R^2Y:

$$R^2Y = 1 - RSS/SSY_{tot.corr.} \qquad \text{(eqn. 4.20)}$$

$$Q^2Y = 1 - PRESS/SSY_{tot.corr.} \qquad \text{(eqn. 4.21).}$$

In the expressions above, $SSY_{tot.corr.}$ represents the total variation in the Y-matrix after mean-centering and scaling.

Without a high R^2Y it is impossible to obtain a high Q^2Y. Generally, a $Q^2Y > 0.5$ is regarded as good and a $Q^2Y > 0.9$ as excellent, but these guidelines are of course heavily application dependent. Differences between R^2Y and Q^2Y larger than 0.2-0.3 indicate the presence of many irrelevant model terms or a few outlying data points.

Furthermore, it should be noted that the R^2Y- and Q^2Y-measures can be equivalently expressed as RSD's (residual SD's) and PRESD's (predictive residual SD's). The latter is often called SDEP (standard deviation of error of prediction). These SD’s should be of similar size to the known or expected “noise” in the system; for example ± 0.3 units for log(1/C) in QSAR investigations.

Regardless of CV mode, "sequential" or "total", the CV-output can be coupled to the ANOVA formalism to accomplish significance testing of derived models [Martens & Martens, 2000; Cederkvist, *et al.*, 2005; Eriksson, *et al.*, 2008]. Thus, CV-ANOVA provides a significance test of the same shape as ordinary ANOVA. The use of cross-validated predictive residuals makes the CV-ANOVA more reliable than ordinary ANOVA. This is particularly important in multivariate PLS and OPLS (see Chapter 5) models where the number of X-variables is often large and diagnostic criteria based on fitted residuals are subjective and risky. Additionally, the CV-ANOVA is faster than permutation tests or bootstrapping based criteria, a highly desirable trait given today's extensive use of small computers in the industrial laboratory environment.

Response permutation tests and other randomization-based test have also been proposed to determine the appropriate number of components in a PLS model [Lindgren, *et al.*, 1996; Rücker, *et al.*, 2007; Wiklund, *et al.*, 2007; Westerhuis, *et al.*, 2008; Xu, *et al.*, 2009]. An illustration and interpretation of the response permutation test as implemented in SIMCA is given in Appendix I.

4.7.4 Standard Errors and Confidence Intervals

Numerous efforts have been made to derive confidence intervals for PLS parameters [see, for example, Denham, 1997]. However, most of these approaches have been based on conventional regression assumptions, treating PLS as a biased regression model. Only in the work of Burnham & MacGregor [Burnham, *et al.*, 1996; Burnham, *et al.*, 1999; Burnham, *et al.*, 2001] has the issue been investigated considering PLS as a latent variable model.

One way to estimate standard errors and confidence intervals directly from the data is to use jack-knifing [Efron & Gong, 1983]. This was actually recommended by Herman Wold in his original PLS work [Wold, H., 1982], and has recently been revived by Martens [Martens & Martens, 2000]. The objective of jack-knifing is to estimate variability of model parameters.

Interestingly, cross-validation – where the objective is to estimate the model complexity giving the optimal predictive power – produces results which can be fed directly to jack-knifing. In this way, the various sub-models generated by cross-validation are used to calculate the standard errors of the model parameters, which are then converted into confidence intervals via the t-distribution. Since the PLS parameters (scores, loadings, etc.) are linear combinations of the original data, they are approximately normally distributed and so jack-knifing works well in the estimation of confidence intervals.

4.8 Use of PLS model: Predictions

When a PLS model is judged reliable – through the interpretation of model parameters, diagnostic tools, cross-validation, etc. – it may then be used for predicting the Y-data for new observations that have not influenced the model. This prediction procedure is illustrated in Figure 4.72, for a situation involving one PLS component.

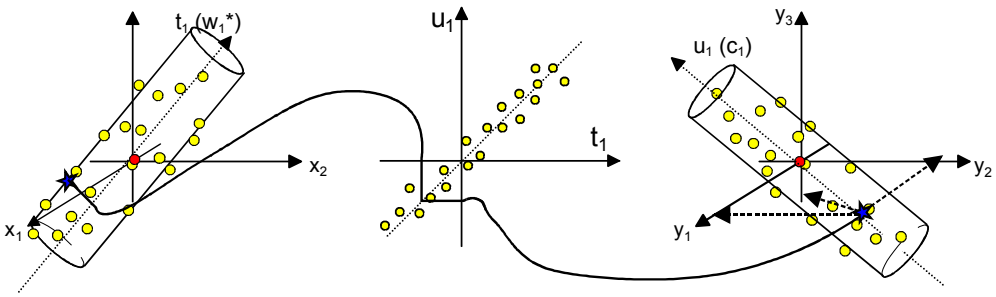


Figure 4.72: The PLS prediction procedure.

A new observation is judged similar to the training set observations if it is located inside the cylindrical tolerance volume in the X-space. Then, its projection on the X-model (t) can be

entered into the t/u inner relation, thus producing a u-value for that dimension. This u-value (or several u-values if there is more than one component) defines a location on the Y-space model, which, in turn, corresponds to a predicted value for each response variable.

Technically, it is of course possible to make PLS predictions also for observations positioned outside the model tolerance volume in the X-space. However, this means that we are extrapolating outside the range of validity of the PLS model. Consequently, these predictions will be much less precise than predictions for observations that fit the model.

It is possible to calculate uncertainties of predicted Y-values in terms of their standard error, or confidence interval. One approach to prediction intervals in PLS was published by Denham [Denham, 1997].

In conclusion, an internally derived estimate of a model's predictive ability can be obtained by cross-validation. However, the only way to be absolutely sure of the predictive power of a model is to make external predictions (see Appendix I), that is, predictions for an independent set of validation observations, and then investigate these observations experimentally.

4.9 Questions for Chapter 4

1. What is a predictor? What is a response?
2. Which three areas are (historically) the major application fields of PLS?
3. Which two operations are commonly used for data pre-processing?
4. What is a t-*score*? What is a u-*score*?
5. What is the inner relation, and how can it be used?
6. What is a PLS-*weight*? How can it be used for model interpretation?
7. Which other variable-related parameters are available for model interpretation?
8. What is a strong outlier? What is a moderate outlier?
9. How can we find out which variables are well explained and which are not?
10. Which tool is used to determine the optimal dimensionality of a PLS model?

4.10 Summary and discussion

4.10.1 Assumptions of PLS

Whenever one wishes to model one or several response variables, Y, by a linear model in correlated X-variables, the PLS method is a good choice [Andersson, 2009]. The graphical representation of the PLS model parameters and residuals facilitates its use and interpretation. The natural assumptions underlying the method, namely that the predictor and response variables are correlated, possibly also noisy and incomplete, are more in line with reality than those of classical regression. In PLS, the variable correlations are modeled as arising from a small set of latent variables. All the measured variables are modeled as linear combinations of these latent variables. In QSAR, for instance, such latent variables are often interpretable in terms of the size, lipophilicity, polarity, polarizability, and hydrogen bonding properties of molecules and their substituents. Hopefully, such an interpretation will also help us understand which factors are regulating the biological response mechanisms.

4.10.2 One or several response variables?

One question which is often debated is whether several Y-variables should be modeled together. PLS *has* the ability to model and analyze several Y-variables together, which has the advantage of giving a simpler picture than separate models for each response. In general, when the Y-variables are strongly correlated, one can recommend that they are analyzed together, since the correlations stabilize the model. If the Y's really measure different things,

however, and hence are fairly independent, one gains little by analyzing them in the same model. On the contrary, with uncorrelated Y-variables the PLS model tends to have many components and is difficult to interpret. The separate modeling of the Y's then gives a set of simpler models with fewer dimensions, which hence are easier to interpret. A reasonable cut-off for "strong" correlation seems to be around 0.5.

To judge whether the Y's are correlated or not it is recommended to start the analysis with a separate principal component analysis (PCA) of the Y-matrix. This will inform about the practical rank of **Y**, i.e., the number of components of the PCA model, A. If A is small compared with the number of Y-variables (M), and if we can understand the resulting components, we can conclude that the Y's are correlated, and a PLS model of all Y's together is warranted. Often, however, one finds from the PCA that the Y's cluster in a few groups according to the nature of activity they measure. This then suggests one separate PLS model for each such group of Y's.

4.10.3 Diagnostic tools

Plots of PLS scores display the relationships among the observations. This is useful, e.g., for finding outliers in the X- or Y-data, and in the relation between **X** and **Y**. In a similar manner, plots of residuals in terms of the DModX and DModY control charts identify moderate outliers in the data. To support decision making, some statistical reference limits are available. For 2-D score plots, Hotelling's T^2 defines a 95% or 99% tolerance region, and for the DModX bar chart, a critical distance corresponding to the 0.05 level is derived from the F-distribution. Observations outside the ellipse defined by Hotelling's T^2 deviate from normality, and observations exceeding the critical distance in DModX, do not fit the model.

The interpretation of a PLS model is based on the weights, w*c, and/or the regression coefficients, b. These weights reflect the relationships among the variables, and tell which are associated and which contribute unique information. Thus, with PLS, one obtains information on what X gives Y, or, how to “set” X to get a desired Y.

4.10.4 Many variables and many observations

Another attractive property of PLS lies in its ability to cope with any shape of data matrix, i.e., a matrix which may have many variables (columns), many observations (rows), or both. The precision and reliability of the PLS parameters related to the observations is enhanced by increasing the number of relevant variables. This property arises because the “new variables”, the scores, are estimated as weighted averages of the X- and Y-variables. Any (weighted) average becomes more precise the more numerical values are used as its basis. Hence, PLS works well with “short and wide” matrices, i.e., matrices with many more columns than rows.

Analogously, the PLS parameters supplying information related to the variables, for example, weights, coefficients, VIP, R^2, Q^2, etc., become more reliable the larger the number of observations. This because these parameters are linear combinations, i.e., weighted averages, of the N observation vectors. Thus, PLS also works well with “long and lean” data tables, with many more rows than columns. Obviously, PLS also handles “square” matrices well, i.e., situations where the number of rows and columns are approximately equal.

4.10.5 Related methods

Before the advent of PLS, two methods were mainly used for regression-like modeling with many and collinear X-variables, namely principal component regression (PCR) and ridge regression (RR). A discussion of PLS, RR, and PCR can be found in the work of Frank and Friedman [Frank and Friedman, 1993]. With PCR, a PCA is first made of the X-matrix, resulting in a score matrix **T** and a loading matrix **P**. In the next step, a few of the first score vectors are used as regressors in a multiple linear regression with the Y-data. Where the first few components of PCA really summarize most of the information in **X** related to **Y**, PCR works nearly as well as PLS [DeJong, 1993]. This is often the case with spectroscopic data, but less common in complicated applications like QSAR.

Ridge regression uses another approach to cope with the near singularities of the X-matrix in the regression problem. Here a small number, d, is added to all the diagonal elements of the variance-covariance matrix of **X**, i.e., **X´X**, before its inversion in the regression algorithm. This is closely related to the omission of all principal components with singular values smaller than d, and indeed RR and PCR demonstrate similar performance. RR can be seen as a circumvention of the multicollinearity problem rather than a solution to it.

Methods such as neural networks (NN) are also often employed in regression analysis. The NN's are equivalent to certain non-linear regressions, and hence are unsuitable for problems involving many and collinear X- and Y-variables. Either a pre-selection of variables or an initial PCA is required for such problems. Examples of NN's in multivariate calibration can be found in the Ph.D. thesis of Greger Andersson [Andersson, G., 1998].

The way in which PLS fits into a framework among other latent variable models has been discussed by Burnham and co-authors [Burnham, *et al.*, 1996; Burnham, *et al.*, 1999]. And, the way in which conventional PLS relates to canonical PLS and continuum power regression, was reported by De Jong and co-workers [De Jong, *et al.*, 2001].

4.10.6 Extensions of PLS

Extensions of the projection methods address more complex problem types, e.g., (i) non-linear PLS [Höskuldsson, 1998; Wold, *et al.*, 1989b], (ii) PLS-discriminant analysis (PLS-DA) [Sjöström, *et al.*, 1986; Ståhle and Wold, 1987], (iii) hierarchical PCA and PLS [Berglund, *et al*., 1997b; Wold, *et al.*, 1996], and (iv) orthogonal PLS (OPLS) [Trygg and Wold, 2002].

A non-linear regression problem may be modeled with PLS by making the $\mathbf{t}_a/\mathbf{u}_a$ inner relation non-linear. For mild non-linearities one can simply expand the X-matrix with the squares and cubes of its columns, and then use this augmented matrix to model **Y**. For instance, in QSAR, non-linear modeling might be relevant when the series of compounds shows large chemical and biological variation, e.g., a biological response variable spanning several orders of magnitude. Chapter 20 deals with non-linear PLS regression.

The objective of PLS-DA is to develop a model that is capable of separating "tight" classes of observations on the basis of their X-variables. The **Y** matrix encodes class membership by a set of "dummy" variables. This "dummy" matrix has *G* columns (for *G* classes) with ones and zeroes, such that the g.th column is one and the others zero for observations of class *g*. A PLS model is fitted between **X** and the artificial **Y**, and in this way a discriminant plane is found in which the observations are well separated according to their class membership. This approach works well provided that each class is "tight" and occupies only a small portion in the X-space. The number of classes must also be fairly low, not more than 5 or 6. PLS-DA is addressed in Chapter 10.

Hierarchical PLS and PCA are two emerging techniques, which simplify the interpretation of very many variables [Wold, *et al.*, 1996; Janné, *et al.*, 2001; Lindström, *et al.*, 2009]. In such a situation, plots and lists of loadings, weights, and coefficients tend to become messy and the results are often difficult to overview. Instead of reducing the number of variables, and thus reducing the validity of the model, a better alternative is often to divide the variables into conceptually meaningful blocks and apply hierarchical PCA or PLS. With hierarchical modeling, each block of variables is first summarized by a few scores. Next, these scores are collected and put together in a new matrix, consisting only of the scores. This new matrix is then modeled with PCA or PLS. Hierarchical modeling is outlined in Chapter 19. Moreover, a special variant of hierarchical modeling tailored for batch processes is described in Chapter 18.

Orthogonal PLS, OPLS, is a recent modification of the PLS method [Trygg and Wold, 2002]. The main idea of OPLS is to separate the systematic variation in X into two parts, one that is related to Y and one that is unrelated (orthogonal) to Y. This partitioning of the X-data will facilitate model interpretation and model execution on new samples. Chapter 5 addresses OPLS in more detail and Chapter 6 O2PLS.

5 Orthogonal PLS (OPLS)

5.1 Objective

The objective of this chapter is to introduce orthogonal PLS (OPLS). OPLS separates the systematic variation in X into two parts, one part that is correlated (predictive) to Y and one part that is uncorrelated (orthogonal) to Y. The result is a model with improved interpretability. To this end, we will use two example datasets. The first is the BeerNIR dataset. It is a tutorial single-Y example covering 50 different beer samples and was performed to investigate the potential of NIR analysis in beer. The second multi-Y example deals with liquid mixtures of four metal-ion complexes. Fifty-two mixtures were formulated according to an experimental design and so the concentration matrix (= Y) contains four variables. To characterize the mixtures their UV-VIS spectra (= X) in the range 310-800nm were measured. OPLS will enable the estimation of the pure profiles of the constituents.

5.2 Introduction

In the foregoing chapter we gave a detailed account of PLS in terms of a geometric perspective and the algebraic expressions involved. We also carefully went through the central PLS plots, what they mean and how they can be interpreted. The intent with the current chapter is to describe an alternative to PLS that is called OPLS. OPLS has certain model interpretation properties rendering the method more tractable than PLS; these properties will be highlighted. In so doing, we shall discuss OPLS from theoretical as well as practical perspectives. We will also clearly distinguish between single-Y and multi-Y OPLS models. Much like the presentation of PLS, the current chapter will provide a walk-through of the OPLS method, containing ample documentation of which parameters are available for model interpretation. Although the beam of light is directed onto OPLS, we will provide a comparison with PLS, so that the interpretational advantages of OPLS become clear.

The current chapter is supplemented by two additional chapters summarizing other aspects of the orthogonal PLS methodology. Chapter 6 discusses O2PLS and its use for data integration, and Chapter 11 introduces the discriminant analysis twist denoted OPLS-DA. In both these chapters additional diagnostic options, beyond the ones presented in the current chapter, will be outlined.

Returning to the foregoing chapter, it was underlined that PLS is a method of dual objectives; it is a projection of X that both (i) approximates X well and (ii) correlates with Y. Formally, the objective function of PLS is one of maximizing the covariance between X and Y. Informally, PLS operates by dividing the variability – the sum of squares – in the X-block in two parts (Figure 5.1), i.e., the systematic and residual parts. The systematic part is the sum of the variability in X that is correlated (predictive) to Y *and* the variability in X that is uncorrelated (orthogonal) to Y. Unfortunately, PLS does not indicate how much of the systematic variability in X is predictive and orthogonal to Y.

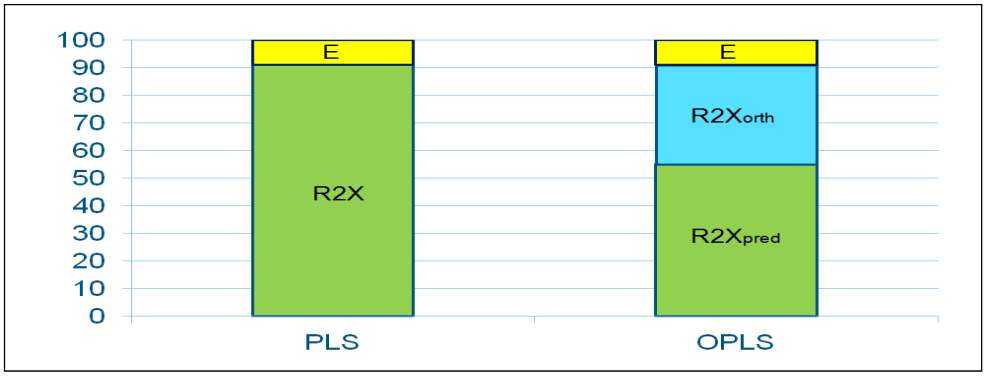


Figure 5.1: PLS (left) divides the variability in the X-matrix in two parts, the systematic variability and the residual variability. OPLS (right) further splits the systematic variability, R2X, in two parts, the part that is correlated (predictive) to Y and the part that is uncorrelated (orthogonal) to Y.

The ability to visualize and interpret the X/Y relationship is a great asset of PLS. However, the complexity of the model interpretation will depend on how the information in X and Y overlap. When a large part of X correlates with a large part of Y, the interpretation of PLS loadings, coefficients, VIPs., etc., is straightforward in the sense that it is indeed the predictive association between X and Y that is reflected in the model parameters. Conversely, however, a much more difficult situation arises when a strong orthogonal structure is present in X. In the latter case, the PLS loadings, coefficients and VIPs will be contaminated by other sources of variations than just those of the true X/Y correlations.

The difficulty in interpreting the regular PLS model increases proportionally to the amount of orthogonal variation present in X. Particularly when there are many components ($A > 3$), any presence of strong orthogonal structure in X will complicate the interpretability. This is because, in regular PLS, the loading weights w_a^* express both the predictive and orthogonal variations in X. The latter is everything varying in X that is not linearly related to Y. This makes w_a^* difficult to interpret directly in later components ($A > 1$).

Formulated in a slightly different way, it can be stated that some of the difficulties in the interpretation of regular PLS parameters stem from the fact that the PLS components are usually not principal components of the X-matrix, i.e. neither the w nor the p loading vectors in PLS are eigenvectors of X´X, but instead w is the eigenvector of X´YY´X. This implies there is no direct score-loading correspondence as in PCA. The interpretation of the relationship between scores (t) and loadings (w, p) in higher-order components is therefore not straightforward.

The OPLS method circumvents the above interpretational problem by separating the predictive and orthogonal variabilities from each other (Figure 5.1). Thus, whereas PLS divides the sum of squares of X in two parts, OPLS divides it in three parts. This yields both good predictions and good interpretability, and in the single-Y case the prediction vector is equivalent to the first principal component of the filtered X-matrix. In other words, in the single-Y case, by theory, the OPLS model, can only have one predictive component [Trygg and Wold, 2002]. Should a single-Y OPLS model comprise more than one component, all components beyond the first one reflect orthogonal variation. However, with multiple Y-variables there can be more than one predictive OPLS component.

In the following sections, the OPLS method will be explained and exemplified. First, we shall discuss an example giving an intuitive understanding of the concepts of predictive and orthogonal variations. Secondly, the OPLS method is outlined from a theoretical perspective. Finally, the ability of OPLS to adequately handle multi-Y calibration problems will be demonstrated. By means of the second example, the estimation of pure profiles of mixture constituents will be considered.

5.3 An example with predictive and orthogonal variations

5.3.1 The BeerNIR dataset

In the food industry, NIR spectroscopy is often applied for the analysis of bulk properties such as total moisture, protein-, fat- and alcohol content. The advantage of the spectroscopic analyses is that they are non-destructive and fast. Potentially an NIR sensor can be implemented online and thereby facilitate real-time quality control. The current example relates to the brewing industry and an attempt to devise a soft sensor for the monitoring of alcohol content in beer, a critical quality parameter that is analyzed on a routine basis in the analytical laboratory. Soft sensors are based on a technique denoted multivariate calibration (see Chapter 8 for an introduction to multivariate calibration).

The BeerNIR dataset is a tutorial example and more information on the beer samples can be found in the literature [Christensen, *et al.*, 2005]. It represents a feasibility study covering 50 different beer samples, and was performed to investigate the potential of NIR in the analysis of beer. NIR-VIS transmission spectra were measured in the range 400-2498 nm, with signal intensities registered at every second nm, thus giving 1050 X-variables. The response variable (the Y-variable) was the alcohol content in the beer. Additional information related to color, pH and bitterness was also measured. Prior to the multivariate data analysis all X-variables were mean centered but not scaled. A discussion on different ways to scale data is given in Chapter 13.

5.3.2 An overview PCA model

PCA was used to accomplish an overview of the 50 (beers) by 1050 (wavelengths) X-matrix. According to default cross-validation configurations, as many as ten principal components are predictively significant. The first two principal components (PCs) dominate the model and account for 98.6% of the spectral variance. The first PC explains 91.2% and the second 7.4%. The scores of these two components are plotted in Figure 5.2.

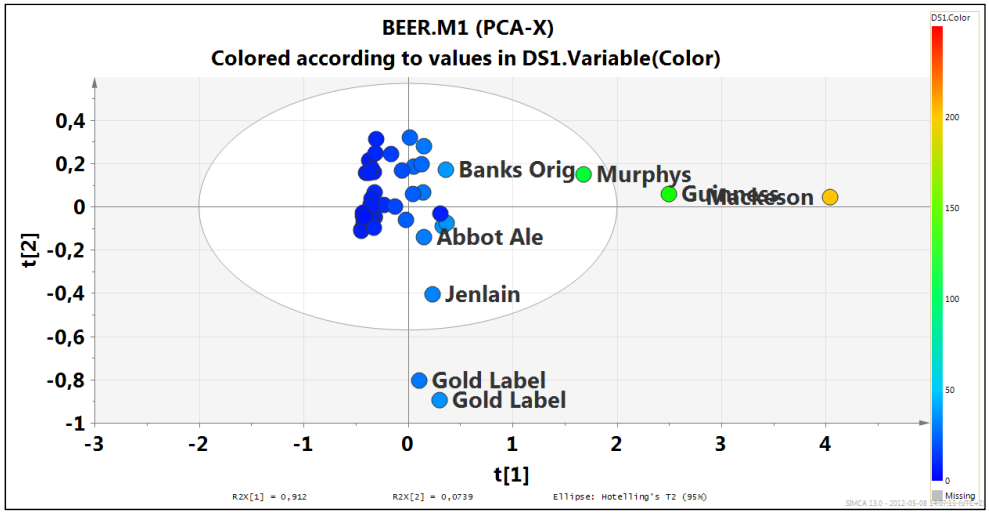


Figure 5.2: Scatter plot of the score vectors, t_1 and t_2, of the first two principal components of the model on the BeerNIR dataset. Each point corresponds to one beer. The separation of the beers in the horizontal direction is governed by their color. The separation of the beers in the vertical direction is influenced by their alcohol content.

As shown by Figure 5.2, the distribution of the beers in the score space is not particularly uniform. This is *not* an optimal starting point for selecting a training set for developing a multivariate calibration model (soft sensor model). Ideally, the dataset should be expanded by including more beer varieties to fill some gaps and accomplish a more balanced distribution of the observations. However, as this is a tutorial example in which the objective is to understand different sources of variations in spectral data, we will proceed by examining the entire dataset.

Interpreting the meaning of a PCA score plot is often facilitated by color coding the observations (beers). In this case, a color-coding is constructed according to the beer color information. Additionally, for reasons of clarity, only a subset of the points have been annotated by their identifiers (name). It turns out that the distribution of the beers in the horizontal direction is regulated by their color, and the darkest ones are situated to the far right in the score plot. Going from left to right among the annotated samples, the numerical value of color increases as 38 (Banks Original), 111 (Murphys), 130 (Guinness) and 202 (Mackeson). This systematic signature from the color is very strong and makes up 91.2% of the variance in the spectral data.

Furthermore, the distribution of the observations along the second PC (the vertical direction) is driven by the alcohol content. This is a weaker signature than color, but still represents a non-negligible 7.4% of the spectral variance. Moving from top to bottom among the labeled observations, the percentage of ethanol increases as 3.4 (Banks Original), 5.0 (Abbot Ale), 6.8 (Jenlain), 9.6 (Gold Label) and 9.8 (Gold Label).

By inspecting the corresponding loading spectra we may confirm our interpretation of the score plot. Figure 5.3 shows the loading spectrum (loading line plot) of p1, the loading vector of the first PC. Evidently, the darkest beers have much higher signal intensities in the visible light region between 400-700nm. That this is indeed the case can be corroborated by plots of the raw spectra (no plots shown).

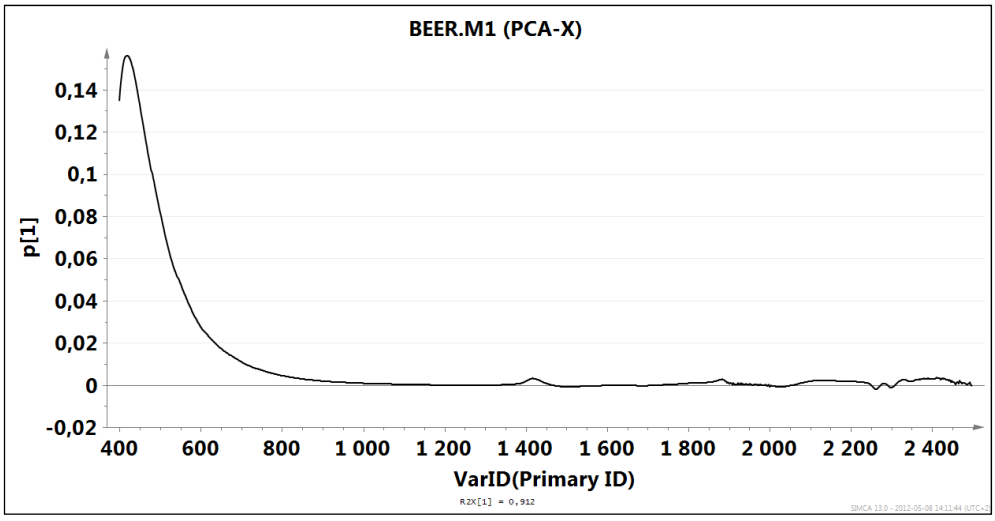


Figure 5.3: Loading line plot of the first loading vector p_1 in the wavelength interval 400-2498nm. The darkest beers have much stronger signal intensities than the lighter beers in the wavelength region 400-700nm (the visible light region).

Analogously to Figure 5.3, Figure 5.4 shows the loading spectrum of the second PC. There are five wavelength regions which characterize the alcohol content in the beers, i.e., the spectral windows of 1380-1510, 1860-1900, 1980-2100, 2240-2320 and 2350-2450 nm.

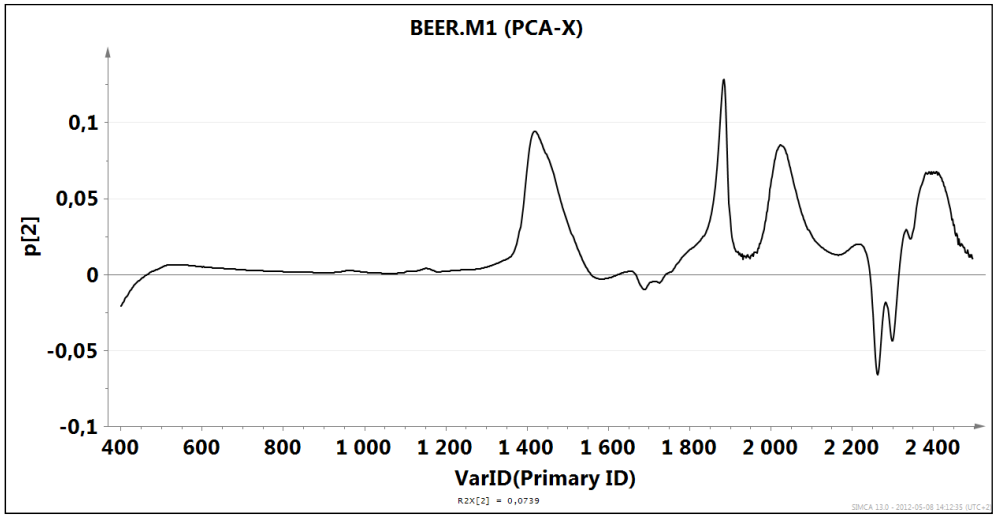


Figure 5.4: Loading line plot of the second loading vector p_2 in the wavelength interval 400-2498nm. The second PC is dominated by variability in the regions 1380-1510, 1860-1900, 1980-2100, 2240-2320 and 2350-2450 nm.

In order to aid the interpretation of the second PC, we have added the spectrum of pure ethanol to the line plot of the second loading vector (Figure 5.5). Note that both p_2 and the ethanol spectrum have been re-scaled to a 0-1 range in order to be plottable in the same graph. Apparently, the double-peak pointing downwards around 2240-2320 nm is directly linked to changes in the ethanol level. The negative loadings of this peak are logical since the beers of the highest ethanol percentages are located at the negative end of the second score vector (cf. Figure 5.2). In fact, this means that the strongest beers have the highest numerical values in this ethanol related split-peak.

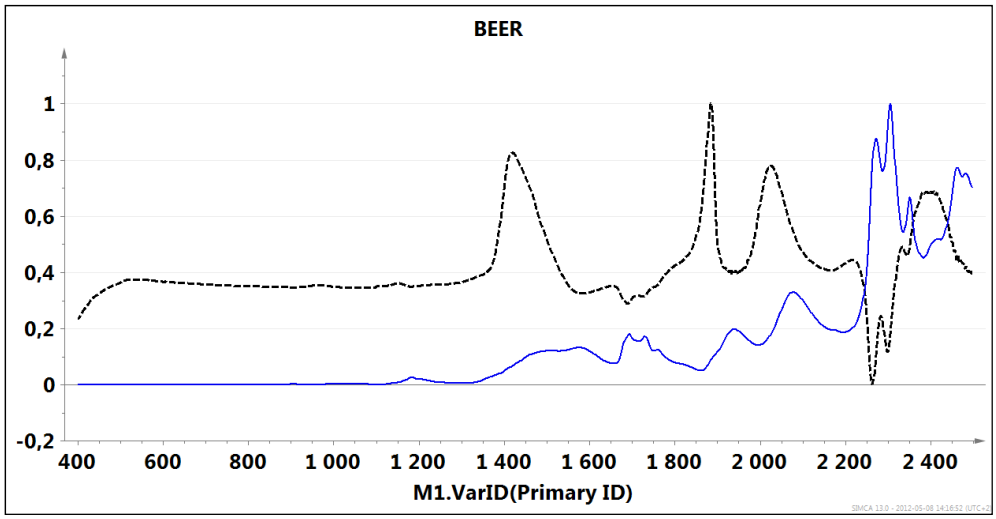


Figure 5.5: Loading spectrum of p_2 together with the pure spectrum of ethanol. The spectra have been re-scaled to conveniently fit into the same plot. The solid line is the pure spectrum of ethanol and the dashed line is p2.

The main conclusion of the PCA model is that the BeerNIR dataset is dominated by two spectral features, i.e., color and alcohol content. The former is the stronger signal. In the next section, we shall investigate how such a dominant, but undesired, spectral feature will influence multivariate calibration modeling, where the primary objective is to define a soft sensor model for alcohol content.

5.3.3 PLS and OPLS models for alcohol content

A PLS model with alcohol content as the single response gave a five-component model with an R2 of 0.99 and a Q2 of 0.98. The model window is displayed in Figure 5.6.

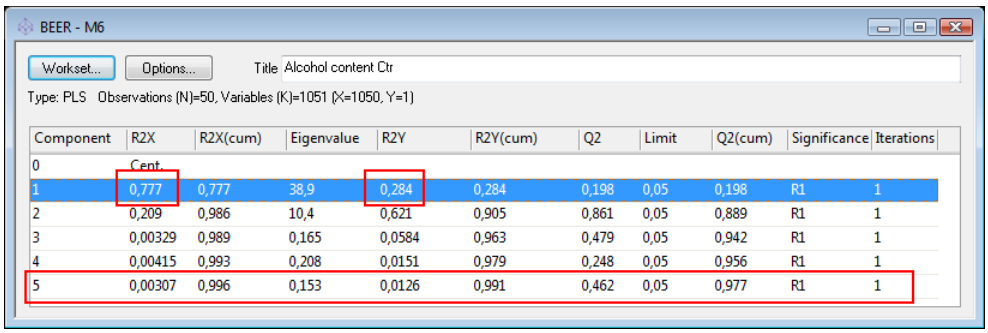


Figure 5.6: Model window of the PLS model. The first component accounts for less than 30% of the variation in Y.

The high R2 and Q2 values reveal that the NIR spectra certainly are related to the alcohol content. It is noteworthy, however, that the first component explains less than 30% of the variation in Y (R2 = 0.28 / Q2 = 0.20). This is due to the fact that the first component is affected by a strong and systematic structure in X (R2X = 0.78) that is not related to the alcohol content, i.e., the variation caused by color.

The analogous OPLS model also has five components. The model window is displayed in Figure 5.7. Using the OPLS notation, the obtained model is a 1+4 model, designating 1 predictive component and 4 orthogonal components. We recall from Section 5.2 that with a single Y-variable the OPLS model can only have one predictive component, but optionally many orthogonal components. All these components can be interpreted in the usual way by inspecting plots of scores and loadings. Experience shows that interpreting the orthogonal components will enhance the total understanding of the data analytical problem at-hand.

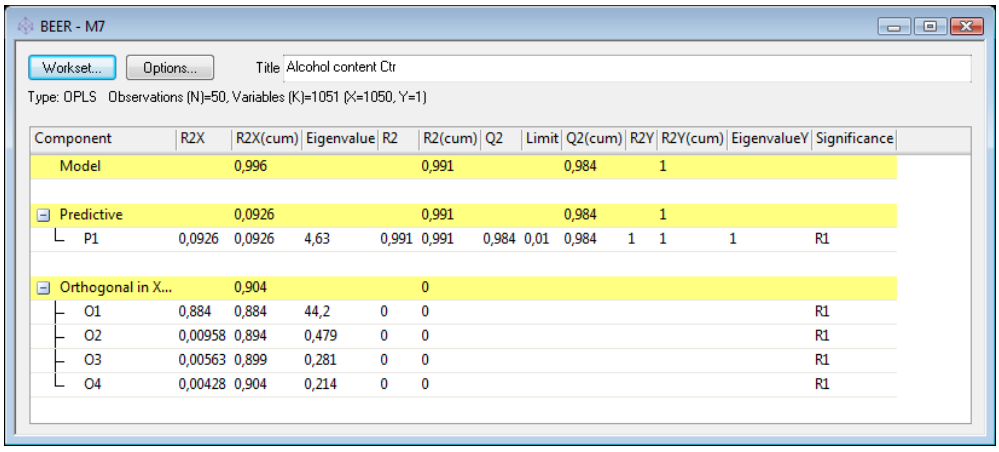


Figure 5.7: Model window of the OPLS model. This model has identical fit and external predictive power to the PLS model. Interestingly, merely 9.3% of the spectral variation is predictive to the alcohol content.

Referring back to Figure 5.1 and the capability of OPLS to divide the variation in the X-matrix in three parts, we realize that the total R2X of the model (= 0.996) is split into the parts R2Xpredictive = 9.26% and R2Xorthogonal = 90.4%. The leftovers, the residuals, amount to 0.04%. Thus, there is nearly a 1:10 relationship between the predictive and orthogonal variations. In the PLS model, this large orthogonal variation masks the predictive variation, whereas with OPLS we are able to "see through" the undesired variability and focus the first component on the variability due to alcohol content. Yet another interesting observation is that the explained variation of the second principal component (see foregoing section) of 7.4% is of similar magnitude to the predictive variation.

Figure 5.8 shows the first score plot of the PLS and OPLS models tiled close to each other. In both cases, the distribution of the points show strong resemblance to the PCA score plot displayed in Figure 5.2. Two dominant directions, corresponding to color and alcohol content, are found. As shown by Figure 5.8, the predictive component of the OPLS model is well aligned with the alcohol content, whereas the corresponding structure is not accounted

for in the first component of the PLS model. This shows the inclination of PLS to pick up strong systematic structures in X in the first component that are not necessarily fully relevant for the Y-variable(s) under study.

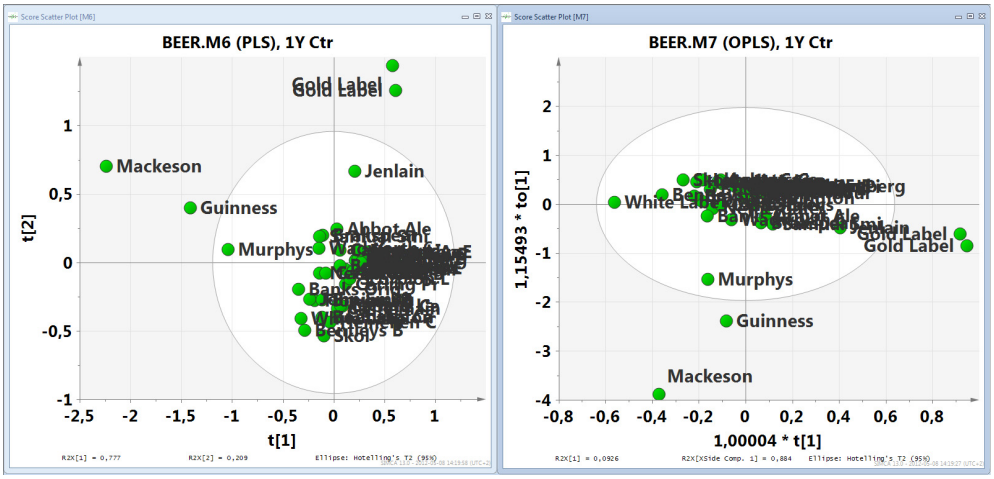


Figure 5.8: Score plots of the PLS (left) and OPLS (right) models. The two models have picked up the two dominant structures in the spectral data, color and alcohol content, but in the OPLS model the predictively relevant information is concentrated into the first component.

The added advantage of separating the predictive and orthogonal variations into separate components is that the analyst gets a better understanding of what is captured by the various components. Figure 5.9 displays the loading spectra of the first two components of the OPLS model, i.e., the predictive component and the first orthogonal component. A comparison of these with the PCA loadings (Figures 5.3 and 5.4) indicates that the second PCA loading spectrum, capturing alcohol content, is very similar to the loading spectrum of the predictive component of the OPLS model, albeit with an inverse sign. There are some differences in the visible region (400-700 nm), but above 700 nm there is an almost perfect inverse relationship between the two loading spectra. A linear regression run in the wavelength region 702-2498 nm region has an R2 > 0.995. Hence, we have ascertained that the predictive component does indeed reflect the alcohol content in the beer varieties.

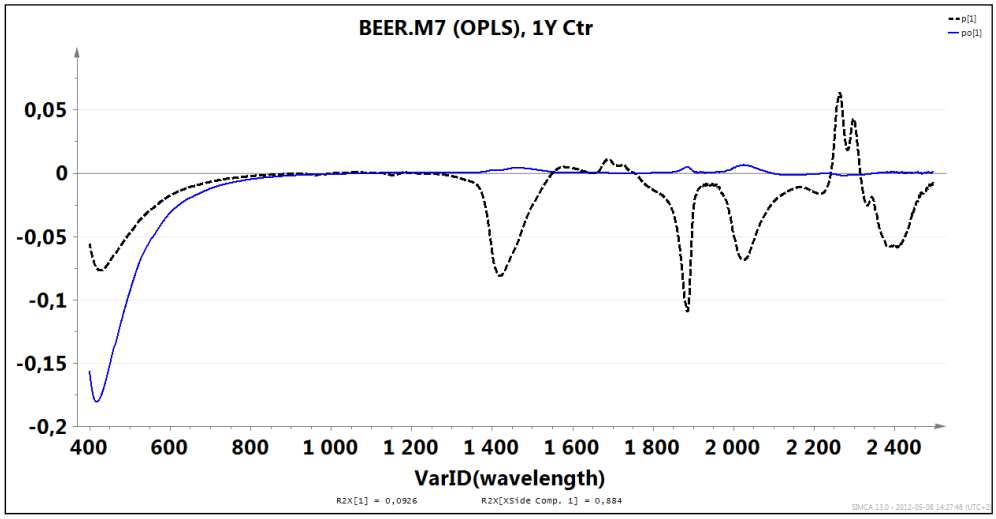


Figure 5.9: Loading spectra of the predictive and the first orthogonal components of the OPLS model on the alcohol content. The dashed line represents the loading of the predictive component and the solid line the loading of the first orthogonal component.

Furthermore, it should be stressed that in the single-Y case, and for the same total number of components, there is absolute identity between a PLS and an OPLS solution as far as fit and external predictivity are concerned. Figure 5.10 shows the relationship between observed and calculated alcohol content for these two models and as evidenced from the inserted

regression lines the R2 is the same in both cases. Hence, there can be no claims that either method predict better or worse than the other! The only difference between PLS and OPLS, in the single-Y case, is the interpretational advantage gained with OPLS.

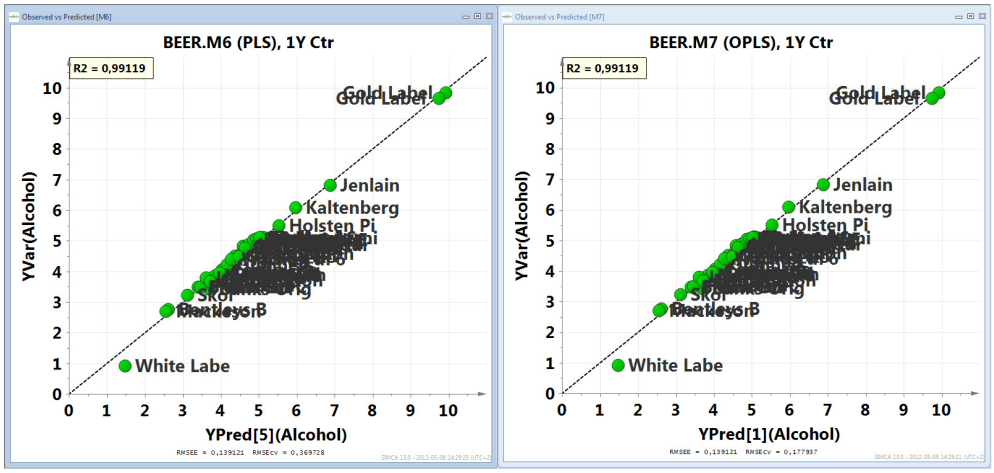


Figure 5.10: Scatter plots showing the relationship between observed and calculated alcohol content for the PLS (left) and OPLS (right) models. The two models have identical fit and external predictivity. PLS needs five components for the predictions whereas OPLS needs the single predictive component to accomplish these results.

This concludes the treatment of the BeerNIR dataset. We have used it to demonstrate the interpretational advantage of OPLS, namely its capability of separating the predictive and orthogonal variations into separate components. This means that the analyst gets a more comprehensive view of what is being interpreted when inspecting the loadings of the various components. In the next few sections, we shall discuss OPLS from a theoretical perspective and illustrate the method using the second multivariate calibration example with multiple Y-variables.

5.4 Theory of OPLS

5.4.1 Objective of OPLS

The objective of OPLS is to divide the systematic variation in the X-block into two model parts, one part which models the correlations between **X** and **Y** and another part which expresses the variation that is not related (orthogonal) to **Y**. Components that are correlated to **Y** are here called *predictive* and are designated by the subscript *P*. Components that are uncorrelated to **Y** are here called *orthogonal* and are designated by the subscript *O*. Note that with a single Y-variable, there is only one predictive component and any number of orthogonal (no relation to **Y**) or pseudo-orthogonal (very little relation to **Y**; only when missing values in **X**) components.

5.4.2 The single-Y OPLS model

In the case of a single response variable, we can write the X-part of the OPLS model as

$$\mathbf{X} = \mathbf{1}\bar{\mathbf{x}}' + \mathbf{t}_p\mathbf{p}_p' + \mathbf{T}_o\mathbf{P}_o' + \mathbf{E} \qquad \text{(eqn. 5.1)}$$

and the OPLS model prediction of y as

$$\mathbf{y} = \bar{\mathbf{y}} + \mathbf{t}_p\mathbf{q}_p' + \mathbf{f} \qquad \text{(eqn. 5.2)}$$

In the expressions above, scores, loadings and residuals have similar depiction to the PLS model, which was outlined in Chapter 4. We re-iterate what was stated in Section 5.4.1,

namely that when the OPLS model encompasses a single Y-variable, only *one* predictive component can result. A consequence of this is that the regression coefficients ($\mathbf{b_P = w_P q'_P}$) become redundant, since the prediction vector is given by the loading p of the predictive component. Similarly, a single-Y OPLS model also makes the VIP plot redundant, as this parameter is directly linked to the loading weight w of the predictive component.

As shown by Figure 5.8 the first scatter plot of a single-Y OPLS model is constructed from plotting the score t_P (just depicted by t in the SIMCA software) of the predictive component together with the score t_O of the first orthogonal component. Note that this score plot is by default scaled proportionally to the R2X of the respective component. Because the concomitant loadings at the same time are normalized to unit length, this leads to a PCA reminiscent score-loading correspondence, which is a desirable property that renders the model interpretation easier. Similar score-loading correspondence also exists on the Y-side of the OPLS model for the scores u and loadings q.

5.4.3 The multi-Y OPLS model

The multi-Y OPLS model can be expressed as follows:

$$\mathbf{X = 1\bar{x}' + T_p P_p' + T_o P_o' + E} \qquad \text{(eqn. 5.3)}$$

and the OPLS model prediction of Y as

$$\mathbf{Y = 1\bar{y}' + T_p Q_p' + F} \qquad \text{(eqn. 5.4)}$$

Thus, in comparison with the single-Y OPLS model, the main difference lies in the fact that there will be more than one predictive component. The number of predictive components is regulated by the number of latent variables in the information overlap between X and Y, which, in turn, often is linked to the rank of Y.

5.4.4 Y-related profiles

When working with spectral data, it is often desirable to estimate the pure spectral profiles as they relate to **Y**. This possibility is exemplified in Section 5.5.7. The pure spectral profiles, termed Y-related profiles in SIMCA, are estimated as:

$$\mathbf{Y\text{-related profiles} = B_P(B_P'B_P)^{-1}} \qquad \text{(eqn 5.5)}$$

where B_P, the OPLS predictive regression coefficients, are computed as:

$$\mathbf{B_P = W_P(P_P'W_P)^{-1}Q_P'} \qquad \text{(eqn 5.6)}$$

Thus, the Y-related profiles are obtained by the transformation of the B_P-matrix into the pure constituent profile estimates. Rotating the coefficients is warranted not only with spectral data, but in any situation, similar to indirect calibration, where the Y-variables are assumed to contribute additively to the variation in **X**, and the Y-matrix is of full rank.

5.4.5 Implementation in SIMCA

OPLS is one of many eligible model types (PCA-X, PCA-Y, PLS-DA, PLS, etc.) in SIMCA. For each computed OPLS model the essential performance statistics are summarized by means of a Model Window. Below, the Model Window of the 1 + 4 OPLS model developed on the BeerNIR dataset is used as an illustration (Figure 5.11).

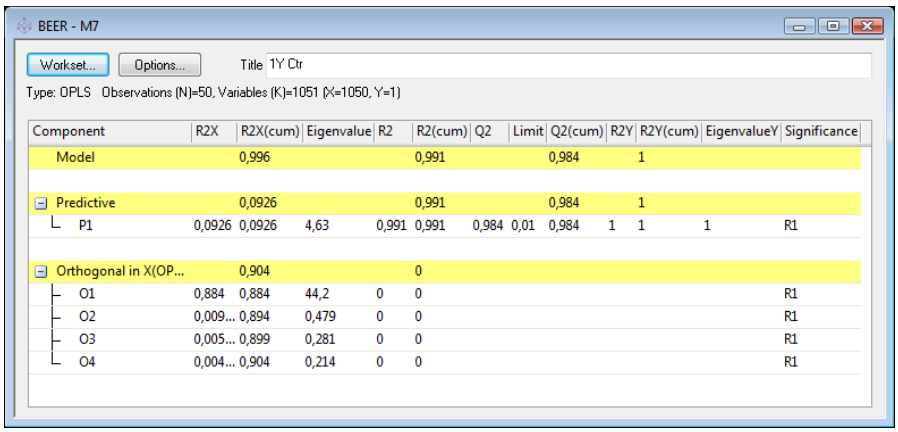


Figure 5.11: Model Overview Window for the 1+ 4 OPLS model developed on the BeerNIR dataset. For explanation, please see the text.

The Model Window in Figure 5.11 displays a summary of the OPLS model. It reveals information on the Predictive and Orthogonal components in X. The Model Window includes:

- a top row summarizing the model, showing the cumulative R2X, R2, Q2, and R2Y;
- predictive components: One row per significant Y-correlated component, showing the predictive components;
- orthogonal components in X: One row per significant Y-orthogonal component, showing the variation in X that is orthogonal to Y.

The table columns are:

- Component - Component index.
- R2X - Fraction of X variation modeled in the component, using the X model.
- R2X(cum) - Cumulative R2X up to the specified component.
- Eigenvalue - The minimum number of observations (N) and variables (K) multiplied by R2X.
- R2 - Fraction of Y variation modeled in the component, using the X model.
- R2(cum) - Cumulative R2 up to the specified component.
- Q2 - Fraction of Y variation predicted according to cross-validation in the component, using the X model.
- Q2(cum) - Cumulative Q2 up to the specified component.
- R2Y - Fraction of the Y variation modeled in the component, using the Y model.
- R2Y(cum) - Cumulative R2Y up to the specified component
- Significance - CV significant or insignificant according to rule.

5.5 A multi-Y calibration study

The second example will be used in two configurations, one with and one without a large portion of orthogonal variation in the X-block. By means of these two cases, a discussion is conducted regarding which parameters should, and should not, be used for model interpretation. It is emphasized that model loadings enable individual components to be interpreted, whereas other parameters are geared to evaluating groups of components, or the performance of an entire model. What is being interpreted is highlighted in the heading of each section.

5.5.1 The Metal-Ion Complexes dataset

The second dataset involves multiple responses and deals with a mixture of four metal-ion complexes. We note that this dataset has previously been analyzed in the literature [Trygg, 2004]. The objective of this example is to highlight the ability of OPLS to estimate the pure spectral profile of known constituents in a mixture as compared to PLS.

Fifty-two (52) mixtures with four different metal-ion complexes were analyzed with a Shimadzu 3101PC UV-VIS spectrophotometer in the wavelength region 310-800 nm, sampling at each wavelength, i.e. the X-matrix comprised 491 X-variables [Trygg, 2004]. The raw spectra are plotted in Figure 5.12.

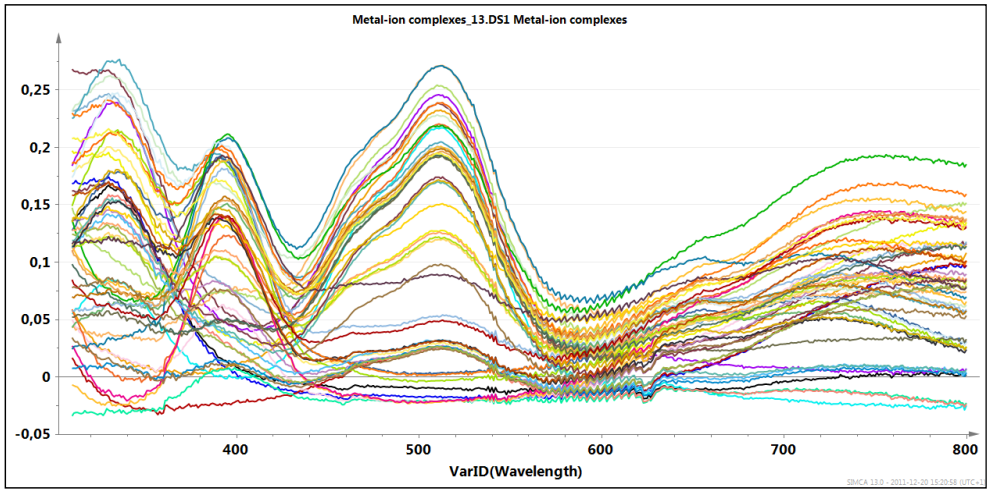


Figure 5.12: Line plot of the unpretreated spectral data (310-800 nm) of the metal-ion complexes.

The metal-ion complexes were mixed according to a design with the following ranges: $FeCl_3$ [0-0.25 mM], $CuSO_4$ [0-10 mM], $CoCl_2$ [0-50 mM], $Ni(NO_3)_2$ [0-50 mM]. Hence, the concentration matrix (the Y-matrix) has four columns. A PCA-Y gives four principal components with an R2 of unity (no plot shown). Hence the rank of Y is four. As a consequence, when modeling the entire Y-block by OPLS, four predictive components are to be expected. A plot of the pure spectral profiles of the metal ions is shown in Figure 5.13.

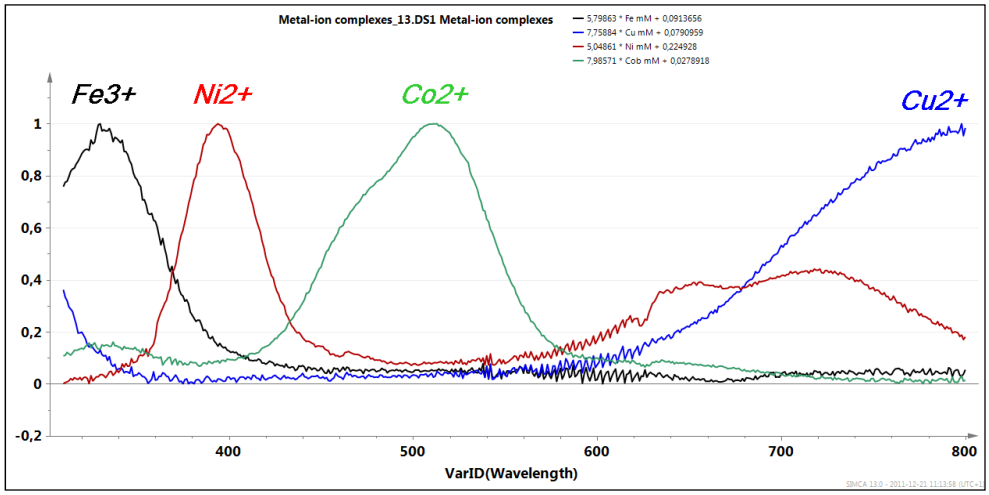


Figure 5.13: The pure spectral profiles of the four metal-ion complexes. All data are scaled to fit the plot well.

5.5.2 OPLS and PLS multi-Y models

The dataset was split into a calibration and a prediction set with 26 observations in each [Trygg, 2004]. Prior to modeling, the spectral data were column centered and the concentration matrix was column centered and scaled to unit variance. PLS and OPLS were then used to model all four Y-variables simultaneously. This corresponds to a situation in

which all spectral variation in the X-matrix is considered known and fairly well controlled, and the proportion of orthogonal variation is small. Hence, it is expected that OPLS and PLS should produce similar models and have similar predictive power for the test set samples.

The OPLS and PLS models fitted to the training set both contained five components. As expected, the OPLS model consisted of four predictive components, and in addition there was one orthogonal component. Figure 5.14 summarizes the performance statistics of the OPLS model. Remember that the cumulative R^2 and Q^2 values are identical for the PLS model (not shown). As shown by Figure 5.14, the major part of the variation in the X-block, 94.7 %, is predictive to Y. Only a minor fraction of the variation in the X-matrix, 5.3%, is orthogonal to Y.

Metal-ion complexes_13 - M3

Workset... Options... Title OPLS training set all Y's

Type: OPLS Observations (N)=26, Variables (K)=495 (X=491, Y=4)

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		1			0,993			0,987		1		
Predictive		0,947			0,993			0,987		1		
P1	0,479	0,479	12,4	0,33	0,33	0,204	0,01	0,204	0,33	0,33	1,32	R1
P2	0,229	0,707	5,94	0,267	0,597	0,159	0,01	0,363	0,272	0,602	1,09	R1
P3	0,175	0,882	4,55	0,173	0,77	0,221	0,01	0,584	0,174	0,776	0,695	R1
P4	0,0646	0,947	1,68	0,223	0,993	0,403	0,01	0,987	0,224	1	0,897	R1
Orthogonal ...		0,0528			0							
O1	0,0528	0,0528	1,37	0	0							R1

Figure 5.14: Summary of fit table of the 4 + 1 OPLS model.

5.5.3 Predictions for the test set

In the next step the external predictive power of the OPLS model was tested by means of the 26 samples in the prediction set. Only results for the 4 + 1 OPLS model are shown (Figures 5.15 – 5.22). Recall that PLS produces similar predictions. It is clear that the predictive power is excellent with external Q^2 values exceeding 0.99 (cf. Figures 5.19 – 5.22).

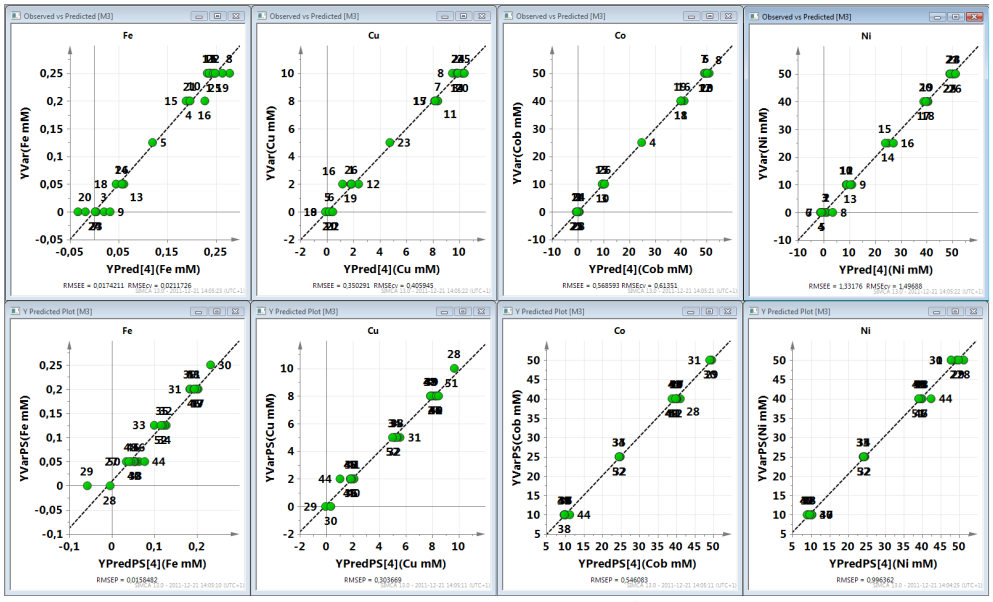


Figure 5.15: (upper left) Relationship between observed and estimated Fe^{3+} concentrations for the calibration set. Figure 5.16: (upper middle left) Same as in Figure 5.15 but for Cu^{2+}. Figure 5.17: (upper middle right) Same as in Figure 5.15 but for Co^{2+}. Figure 5.18: (upper right) Same as in Figure 5.15 but for Ni^{2+}. Figure 5.19: (lower left) Relationship between observed and predicted Fe^{3+} concentrations for the prediction set. Figure 5.20: (lower middle left) Same as in Figure 5.19 but for Cu^{2+}. Figure 5.21: (lower middle right) Same as in Figure 5.19 but for Co^{2+}. Figure 5.22: (lower right) Same as in Figure 5.19 but for Ni^{2+}.

5.5.4 Interpretation of the single orthogonal component of the 4+1 OPLS model

Having verified the external predictivity of the OPLS model, we may proceed by interpreting its meaning. The loading spectrum of the orthogonal component (Figure 5.23) suggests that the orthogonal variability is mainly due to a spectral offset (all loading values are of similar size and sign). Interestingly, there seem to be some spectral artifacts around 460nm and 625nm. Such spectral features could arise from a detector shift or some spectroscopic phenomenon beyond control. This ability of OPLS to flush out non-correlating spectral variation as the orthogonal part of the model is highly interesting, for instance in calibration transfer and time-trend detection.

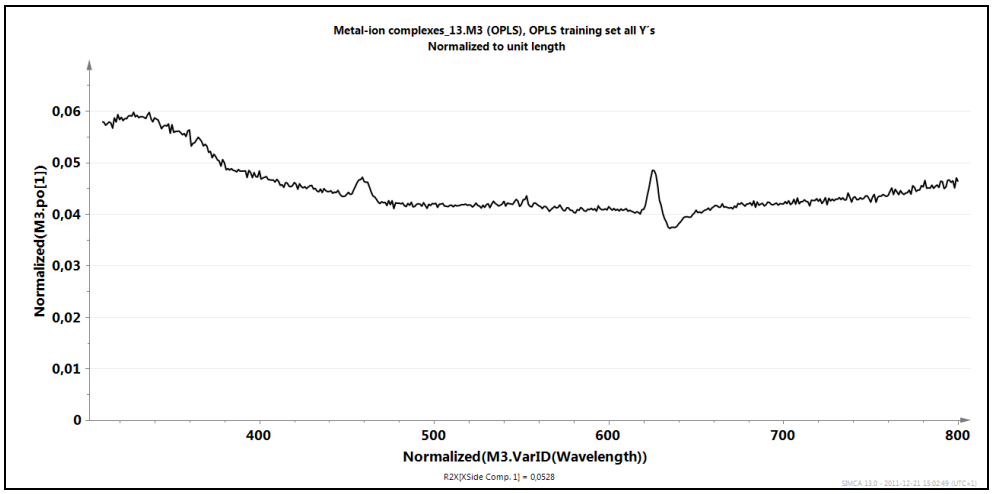


Figure 5.23: Loading profile of the orthogonal component of the 4 + 1 OPLS model. Two spectral artifacts around 460 nm and 625 nm are easily spotted.

5.5.5 Interpretation of the predictive components of the 4+1 OPLS model using loadings

The interpretation of the first predictive component of the OPLS model indicates that it reflects changes in the Co^{2+} concentrations (Figure 5.24). Figure 5.24 shows a line plot of the measured spectrum of Co^{2+} together with the first loading of the OPLS and PLS models. Evidently, p_1 of OPLS and p_1 of PLS are almost identical in this case. The reason for this is that the orthogonal variation in X is small, and hence the PLS solution is little perturbed by features in X uncorrelated to Y.

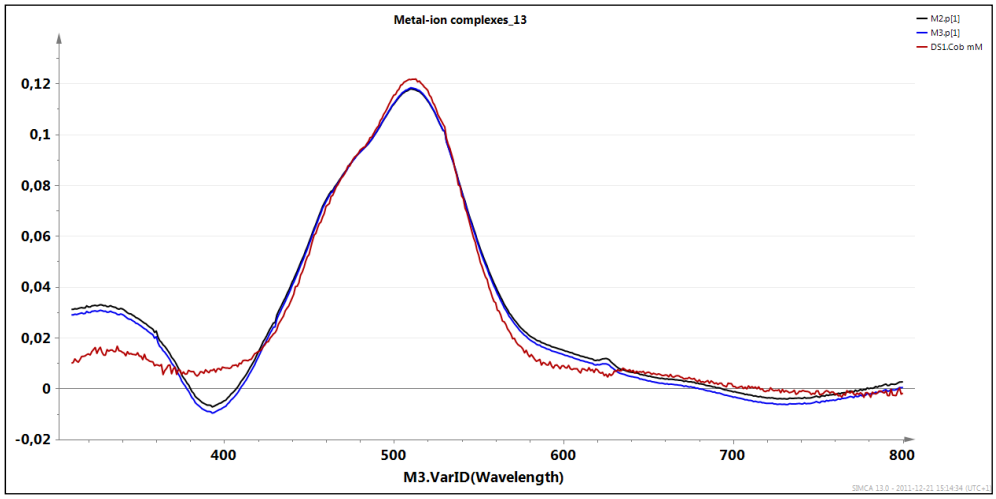


Figure 5.24: Loading profile of the first predictive component of the 4+1 OPLS model (blue) and the first component of the corresponding PLS model (black). The two loadings are almost identical in this case, since the amount of orthogonal variation in X is small. The measured spectrum of Co^{2+} is given in red.

Analogously, the interpretation of the second predictive component of the OPLS model indicates that it represents changes in the Fe^{3+} concentrations. The measured spectrum of Fe^{3+} together with the second loading of the OPLS and PLS models are displayed in Figure 5.25.

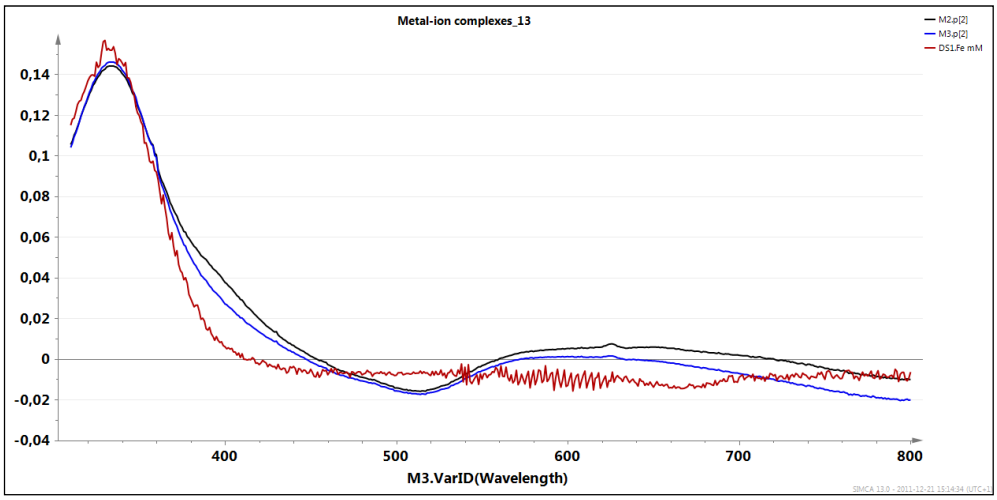


Figure 5.25: Loading profile of the second predictive component of the 4+1 OPLS model (blue) and the second component of the corresponding PLS model (black). The measured spectrum of Fe^{3+} is given in red.

Further model interpretation shows that the changes in Cu^{2+} and Ni^{2+} concentrations are modeled by a combination of the third and fourth predictive components of the OPLS model (no plots given). However, the similarity between the OPLS and PLS loading profiles decreases in the higher order components and for instance p_4 of the PLS solution is inverted w r t p_4 of the OPLS solution (no plots shown).

The conclusion is that when the proportion of orthogonal variation in X is small, using the p loading vector for interpreting PLS and OPLS models is a sound choice.

5.5.6 Interpretation of the 4+1 OPLS model using coefficients

Another alternative for model interpretation is the coefficients. Since the OPLS model covers four Y-variables, there will be four sets of coefficients. Figure 5.26 shows a line plot with the OPLS and PLS coefficients for Fe^{3+} together with the pure spectrum of Fe^{3+}. Note that the coefficients relate to the entire prediction model in both cases. We can observe two things. First of all, the OPLS and PLS coefficients are very similar and the respective lines displayed in Figure 5.26 are almost indistinguishable. This close resemblance arises because the proportion of orthogonal variation in X is small. (In Section 5.6, we shall undertake a modification of the metal-ion dataset leading to very different coefficient profiles of the OPLS and PLS models.) Secondly, the coefficients for Fe^{3+} only partly mimic the pure spectrum of Fe^{3+}. The pure spectrum peak at 310-350 nm is well reproduced by the OPLS and PLS coefficients, but the coefficients also display other peaks that are not related to the pure spectrum structure.

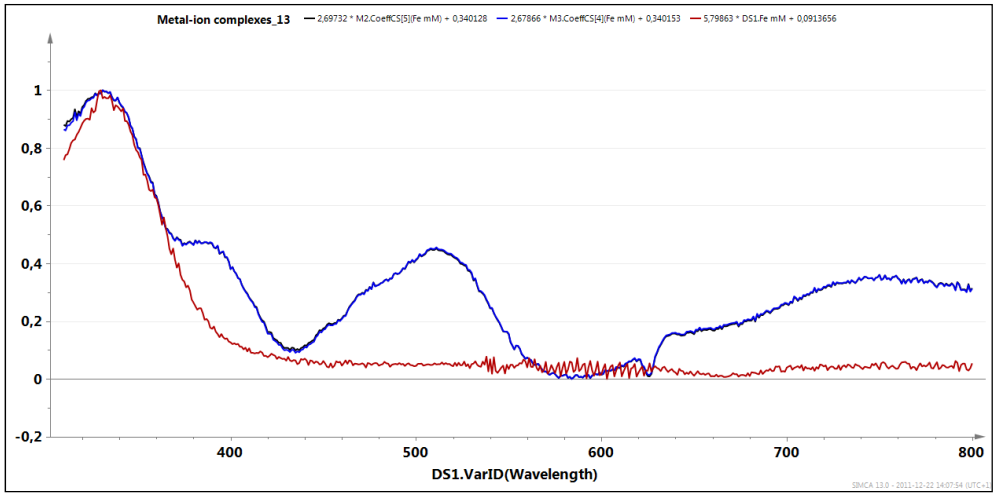


Figure 5.26: Coefficient profiles for Fe^{3+}. The coefficients of the 4+1 OPLS model (blue) and the PLS model (black) are almost indistinguishable. The measured spectrum of Fe^{3+} is given in red. Both sets of coefficients well reproduce the Fe^{3+} peak between 310-350 nm, but also contain other peak-like structures that are not related to the pure spectrum in question. Note that the coefficients and pure spectrum have been re-scaled to a 0-1 representation.

Continuing with the coefficients for Cu^{2+} and the corresponding pure spectrum (Figure 5.27), we can see that the coefficients resemble the pure spectrum in the 650-800 nm region. However, as with the coefficients for Fe^{3+}, there are additional peaks that are not caused by changes in the concentrations of Cu^{2+}.

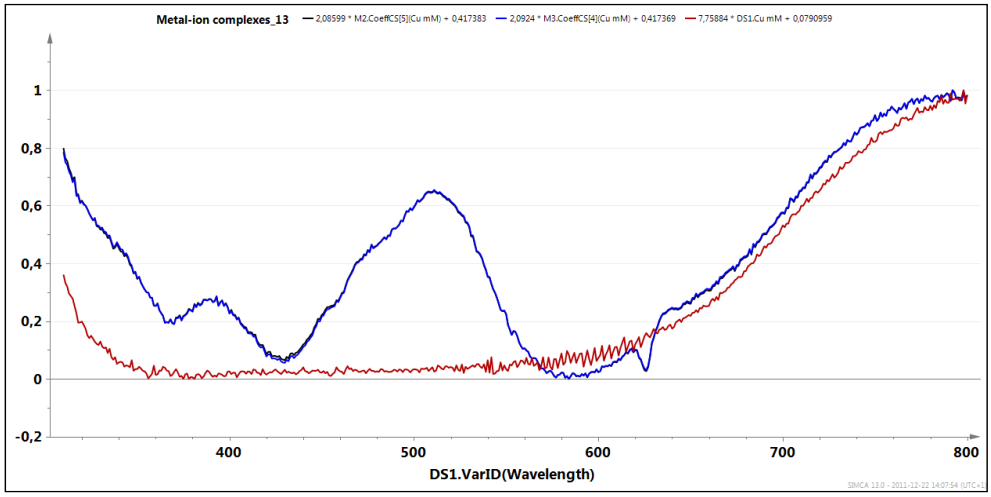


Figure 5.27: Coefficient profiles for Cu^{2+}. The coefficients of the 4+1 OPLS model (blue) and the PLS model (black) overlap. The measured spectrum of Cu^{2+} is given in red. The PLS and OPLS coefficients resemble the pure spectrum in the 650-800 nm region. Note that the coefficients and pure spectrum have been re-scaled to a 0-1 representation.

The situation is similar with regards to the remaining metal ions, Co2+ and Ni2+. The profiles of the OPLS and PLS coefficients are strongly overlapping, but apart from mapping the respective main peaks well, the coefficients indicate the existence of extra peaks that are not real (Figures 5.28 and 5.29).

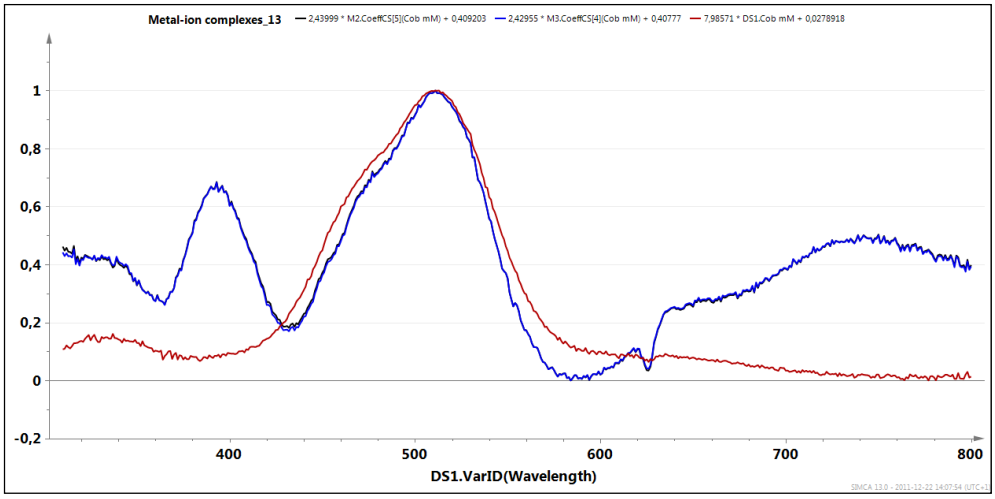


Figure 5.28: Coefficient profiles for Co^{2+}. The coefficients of the 4+1 OPLS model (blue) and the PLS model (black) overlap. The measured spectrum of Co^{2+} is given in red. The PLS and OPLS coefficients resemble the pure spectrum in the 450-550 nm region. Note that the coefficients and pure spectrum have been re-scaled to a 0-1 representation.

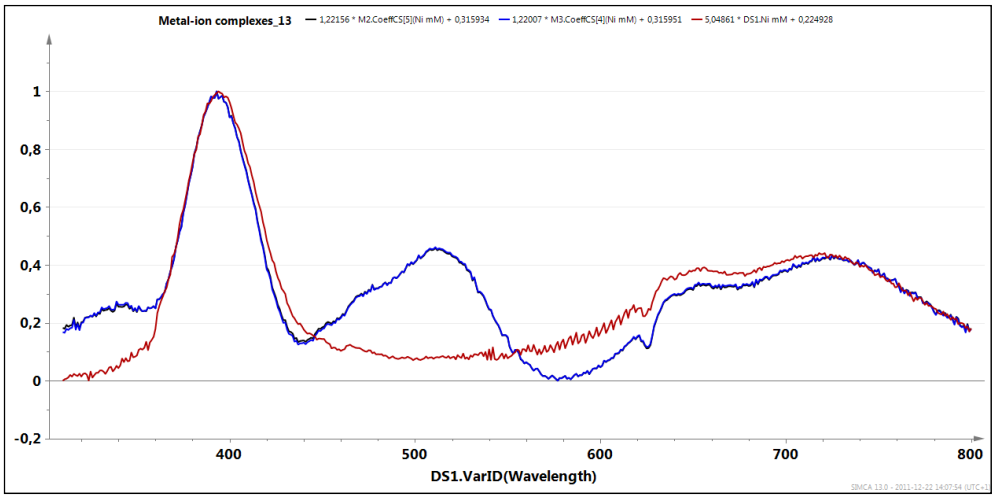


Figure 5.29: Coefficient profiles for Ni^{2+}. The coefficients of the 4+1 OPLS model (blue) and the PLS model (black) overlap. The measured spectrum of Ni^{2+} is given in red. The PLS and OPLS coefficients resemble the pure spectrum in the 360-440 nm region. Note that the coefficients and pure spectrum have been re-scaled to a 0-1 representation.

The conclusion that can be drawn from the inspection of Figures 5.26-5.29 is that the OPLS and PLS coefficients are very similar. The reason for this similarity is the small portion of orthogonal variation in the spectral data. In contrast to the loading p, however, plots of coefficient profiles do not necessarily reveal the true shape of the pure spectra of the metal ions. Fortunately, the coefficient vectors can be rotated into a form, which is directly linked to the constituent pure profiles (see equation 5.5). In the next section, the use of such pure spectrum coefficients is considered.

5.5.7 Interpretation of the predictive part of the 4+1 OPLS model using coefficients rotated into pure spectrum profiles

As demonstrated above, plots of OPLS and PLS coefficients do not properly reflect pure spectra of the constituents in a complex mixture. However, as discussed in Section 5.4.4, the OPLS predictive regression coefficients, B_P, may be re-expressed such that they really do reflect the pure profile of each constituent, the Y-related profiles. A pre-requisite for this re-expression is, however, that the Y-variables can be assumed to contribute additively to the variation in X, and that the Y-matrix is of full rank. Please note that “full rank” is not

equivalent to perfect orthogonality between the columns of Y, i.e., the columns of Y can be partially correlated. Recall that PCA was used to verify that the rank of Y is four, so this re-expression is valid in the current case. The Y-related profiles are plotted in Figure 5.30. For convenience, Figure 5.31 presents the pure spectral profiles that were originally plotted in Figure 5.13. The similarity between the two plots cannot be mistaken.

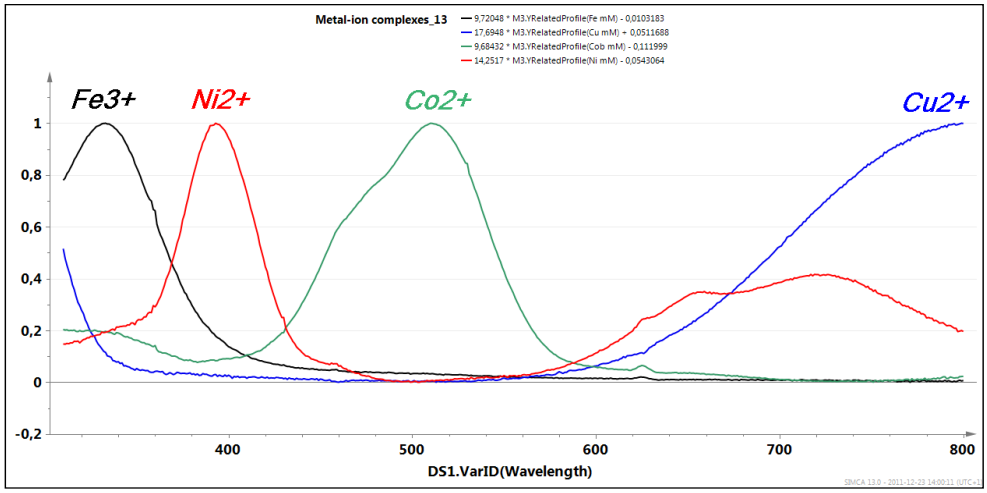


Figure 5.30: The four Y-related profiles. They are estimates of the pure spectral profiles of the four metal-ion complexes. All data are scaled to fit the plot well.

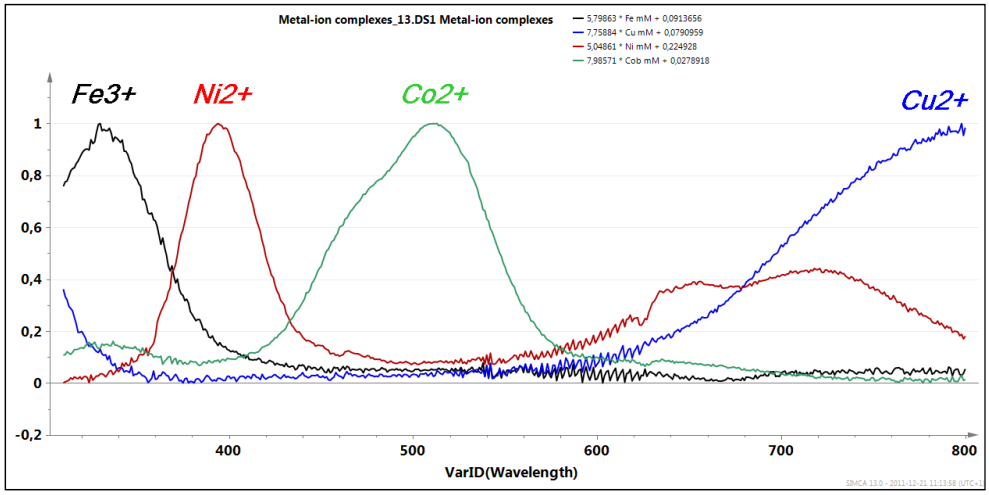


Figure 5.31: The pure spectral profiles of the four metal-ion complexes. All data are scaled to fit the plot well.

A similar rotation was performed for the coefficients of the PLS model, the results of which are displayed in Figure 5.32. It should be noted that the rotation of the coefficients in this case encompasses all five components of the PLS model. The similarity to Figures 5.30 and 5.31 is due to the low proportion of orthogonal variation in the spectral data.

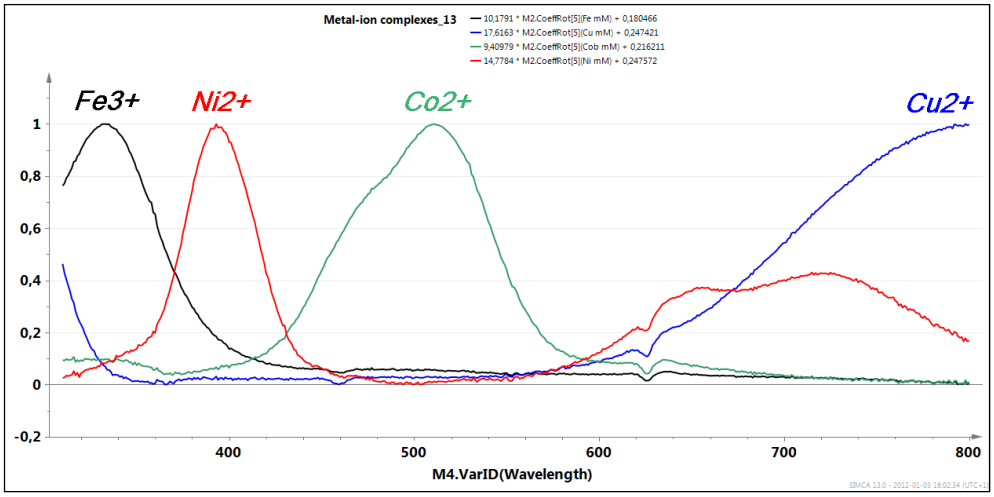


Figure 5.32: The rotated coefficients of the PLS model estimating the pure profiles of the four metal-ion complexes. All data are scaled to fit the plot well.

In summary, as long as the assumptions behind the Y-related profiles are fulfilled, they represent a viable complement to the loadings p for interpreting the predictive components of an OPLS model. Moreover, the estimation of pure spectral profiles also work surprisingly well in the presence of orthogonal variations in the X-block. The latter phenomenon is illustrated in Section 5.6. It should also be noted that in the SIMCA software the use of the Y-related profiles is restricted to OPLS models (these profiles are not available for PLS models).

5.5.8 Interpretation of the 4+1 OPLS model using VIP

The last parameter that we shall consider in this context is the VIP parameter. We recall from Section 4.6.3 that VIP is a parameter which summarizes the importance of the X-variables, both for the X- and Y-models. Its attraction lies in its intrinsic parsimony; for a given model and problem there will always be only one VIP-vector, summarizing all predictive components and Y-variables. By default a VIP plot is size sorted, but VIP can also be displayed as a function of any physical or chemical ordering structure among the variables, i.e., in this case wavelength. Figure 5.33 shows line plots of the VIP vector for the OPLS and PLS models. Given the similarity of the loadings of the two underlying models, the congruence of the two VIP profiles comes as no surprise (cf. Figures 5.24 and 5.25). Due to the small amount of orthogonal variation in the spectral block of data, the two VIP-curves can be thought of as an overlay of the four Y-related profiles rendered in Figure 5.30.

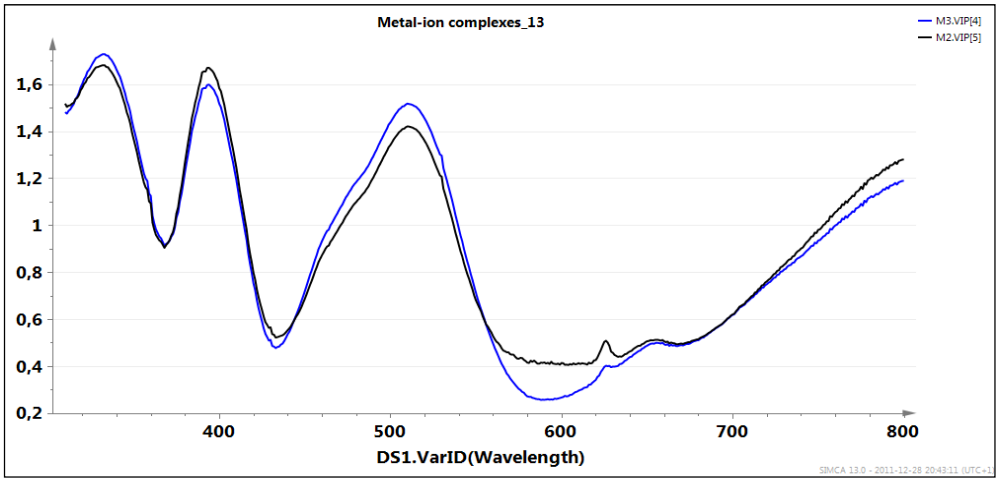


Figure 5.33: VIP profiles for the OPLS and PLS models. The VIP-values of the 4+1 OPLS model (blue) and the PLS model (black) are very similar. In essence, the two congruent VIP-curves can be thought of as an overlay of the four Y-related profiles rendered in Figure 5.30.

5.5.9 Summary; model interpretation in the case of low degree of orthogonal variation

We have now discussed the most common variable related parameters that are available for interpreting the meaning of an OPLS model, using a multivariate calibration application. In the current example, most of the reviewed parameters make sense, either in standalone format or in conjunction with other parameters, except for the use of the coefficients, which display ghost peaks that are not related to the Y-variable in question. In fact, even the use of VIP for interpreting the relationship between the spectral data (X) and the concentration matrix (Y) is warranted and does not lead to incorrect inference. (However, remember that for the single-Y case, VIP is made redundant since the OPLS model will then merely comprise one predictive component.) The bottom line is that as long as the amount of orthogonal variation in X is small, most parameters are easy to interpret. As we shall see in the next section (Section 5.6), however, this reliability is greatly compromised when the amount of orthogonal variation in the spectral data rises to significant proportions. The take-home message from the current example is that a combination of the loading p and the Y-related profiles offers a powerful tool for interpreting a multi-Y OPLS model based on spectral data.

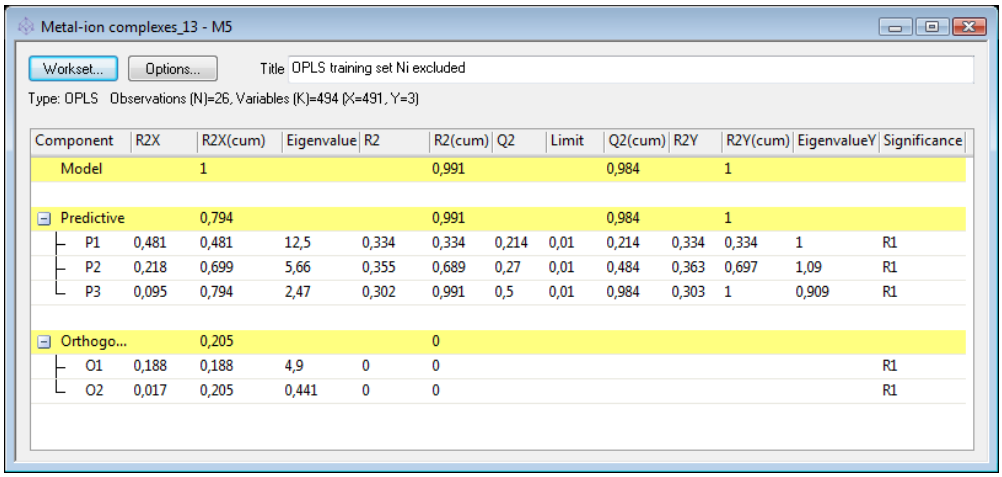
5.6 What happens when the orthogonal variation increases?

The material presented in Section 5.5 demonstrates that when there is little orthogonal variation in the X-block, OPLS and PLS perform similarly. However, with increasing proportions of orthogonal variation, this similarity decreases and eventually we end up in a situation where OPLS is the preferred choice for model interpretation. In order to illustrate the interpretative challenges that arise with increasing proportions of orthogonal variation, we shall artificially induce that kind of situation in the Metal-ion tutorial dataset.

Thus, in this second analysis of the Metal-ion dataset, one of the responses was excluded from the Y-matrix, i.e., the number of Y-variables was reduced from four to three. We choose to eliminate the Ni^{2+}-variable, because from visual inspection (cf. Figure 5.31) the spectral profile of Ni^{2+} has the greatest overlap with the other pure profiles. This corresponds to creating the situation in which not all constituents in X are known in Y, i.e. the X-matrix should contain a large amount of orthogonal variation.

5.6.1 Exclusion of the Ni^{2+}-response from Y (three Y-variables)

The OPLS and PLS models fitted to the calibration set samples again contained five components. This time, however, the OPLS model had a 3 + 2 configuration, see the model summary table in Figure 5.34. We can see that the orthogonal variation in X represents 20.5%. The overall summary statistics of the PLS model are similar to the 3 + 2 OPLS model. The OPLS and PLS predictions for the test set obtained without Ni^{2+} are very similar and so, for conciseness, are not shown.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		1			0,991			0,984		1		
Predictive		0,794			0,991			0,984		1		
P1	0,481	0,481	12,5	0,334	0,334	0,214	0,01	0,214	0,334	0,334	1	R1
P2	0,218	0,699	5,66	0,355	0,689	0,27	0,01	0,484	0,363	0,697	1,09	R1
P3	0,095	0,794	2,47	0,302	0,991	0,5	0,01	0,984	0,303	1	0,909	R1
Orthogo...		0,205			0							
O1	0,188	0,188	4,9	0	0							R1
O2	0,017	0,205	0,441	0	0							R1

Figure 5.34: Summary of fit table of the 3 + 2 OPLS model.

5.6.2 Interpretation of the first orthogonal component of the 3+2 OPLS model

Among the two orthogonal components of the OPLS model, the first one accounts for almost 19% of the spectral variance. The last orthogonal component is of marginal importance, and will not be considered here. The loading vector of the first orthogonal component, p_{1O}, is plotted as a line plot in Figure 5.35 together with the pure spectrum of Ni^{2+}. The two spectra have been rescaled to a 0-1 range to facilitate comparison. Due to their close resemblance, it is inferred that the first orthogonal component accounts for the spectral properties of Ni^{2+}, as expected. The reason for this behavior is that the variability due to Ni^{2+} is expressed in X but at the same time the corresponding information is lacking in Y.

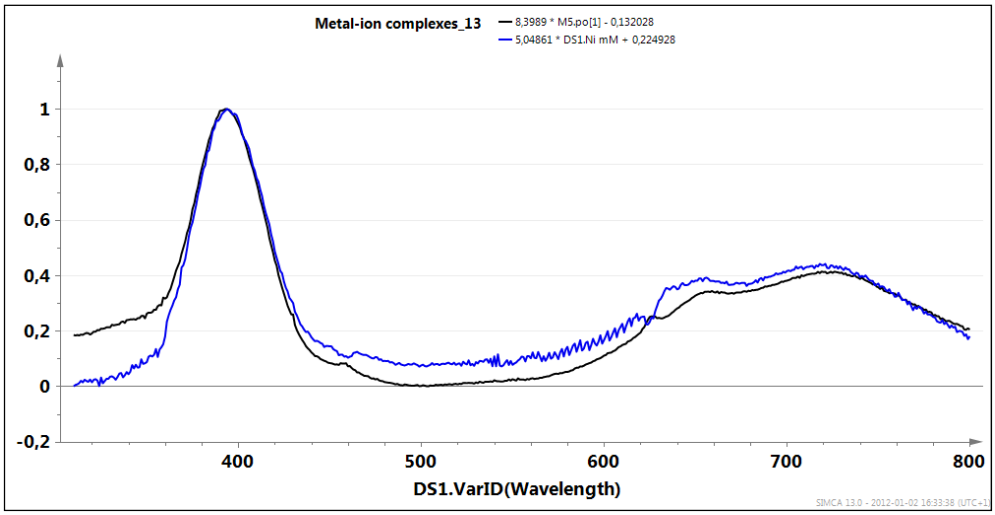


Figure 5.35: Loading profile (black line) of the 4th component (the first orthogonal) of the 3 + 2 OPLS model. The blue line represents the pure spectrum of Ni^{2+}. The two spectra have been rescaled to a 0-1 range.

5.6.3 Interpretation of the predictive components of the 3+2 OPLS model using loadings

Figures 5.36 – 5.38 display the loadings of the three predictive components of the OPLS model. For comparison the loadings of the first three components of the PLS model are given together with some of the pure spectrum profiles. Figure 5.36 indicates that the first predictive OPLS loading, the first PLS loading and the pure spectrum of Co^{2+} are very

similar. As such, this plot is almost a copy of Figure 5.24. This means that the first components of the current OPLS/PLS models capture the same type of information (Co^{2+} variability) as the previous models fitted to four Y-variables.

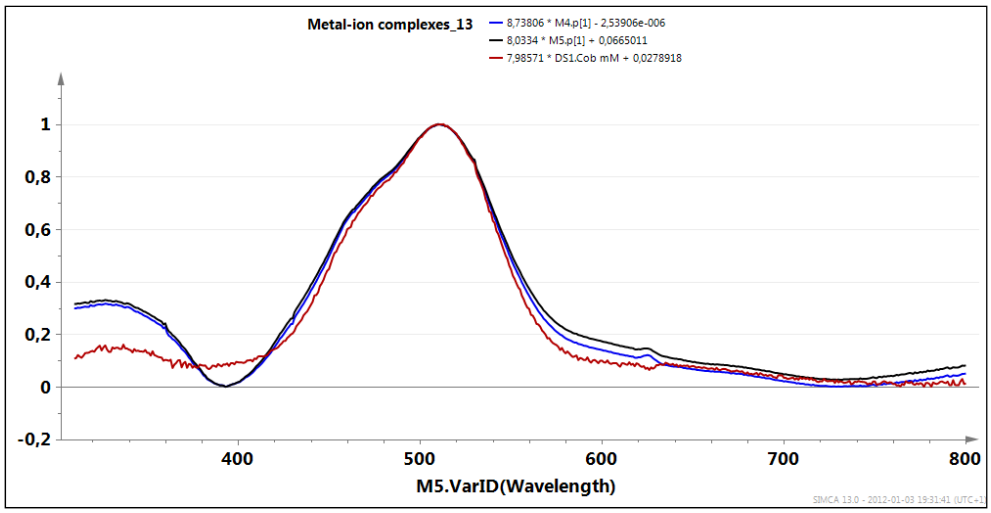


Figure 5.36: Loading profile of the first predictive component of the 3+2 OPLS model (blue) and the first component of the corresponding PLS model (black). The two loadings are similar. The measured spectrum of Co^{2+} is given in red.

Continuing with the second predictive OPLS loading and the second PLS loading we can deduce from Figure 5.37 that they mirror the variability in Fe^{3+} concentrations. Hence, the information in this plot is equivalent to the information obtained from Figure 5.25. This means that the second predictive component of both the 4+1 and the 3+2 OPLS models capture the same phenomenon, the signature of which is little perturbed by the increase in orthogonal variation.

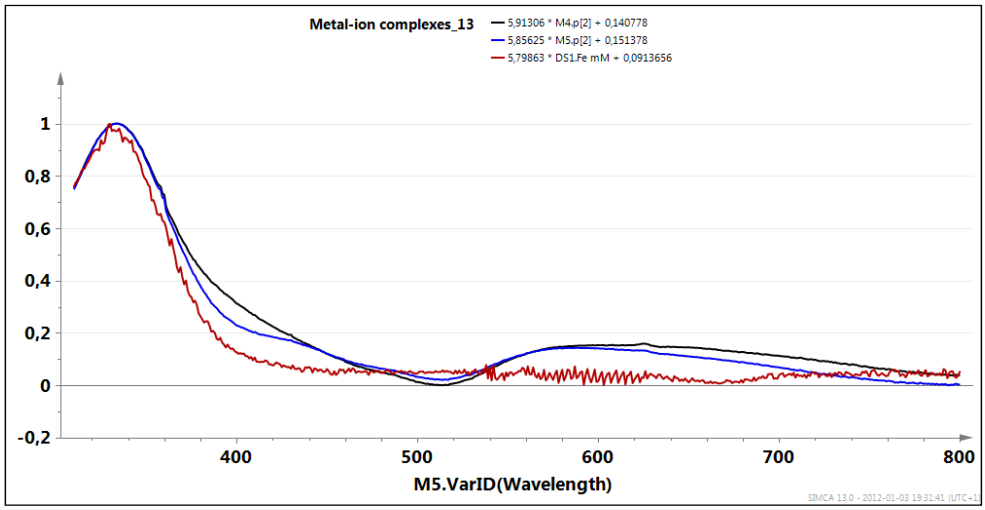


Figure 5.37: Loading profile of the second predictive component of the 3+2 OPLS model (blue) and the second component of the corresponding PLS model (black). The two loadings are similar. The measured spectrum of Fe^{3+} is given in red.

So far, the insight is that the increased orthogonal variability has had little impact on the 4+1 and 3+2 OPLS models, and their corresponding PLS counterparts, at least as far as the two first components of each model are concerned. However, the impact of the orthogonal variability is strongly manifested in the third component of the current PLS model. As shown by Figure 5.38, the third loading of the PLS model "remembers" the structure of Ni^{2+}, although that structure is no longer present in the X/Y overlap. Hence, the PLS model has confounded predictive and orthogonal variation in the third component. Recall the Ni^{2+} peak was accounted for by the first orthogonal component of the 3+2 OPLS model (see

loading profile in Figure 5.35). The third predictive loading of the 3+2 OPLS model is almost devoid of the Ni^{2+} peak. Instead the structure of this loading is more in line with the profile of the pure spectrum of the last modeled constituent, i.e. Cu^{2+}.

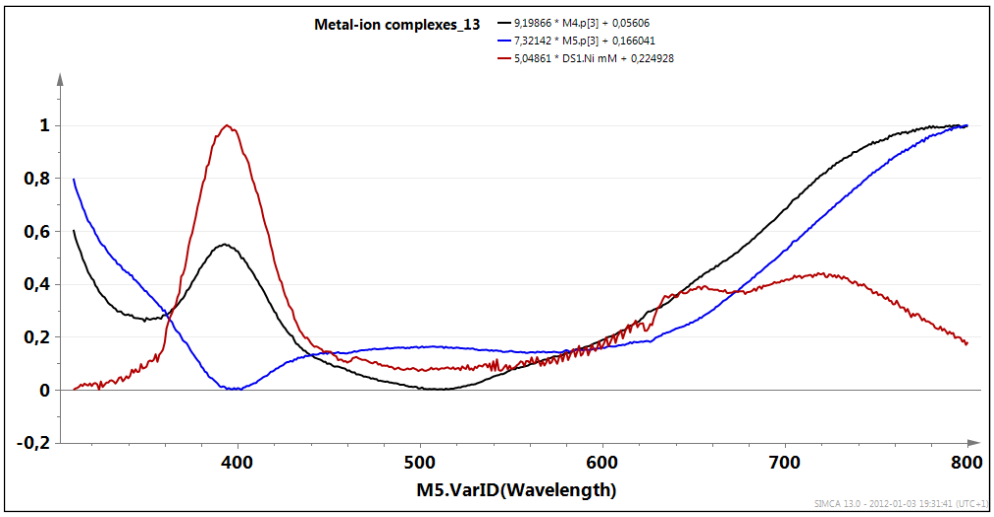


Figure 5.38: Loading profile of the third predictive component of the 3+2 OPLS model (blue) and the third component of the corresponding PLS model (black). The two loadings are very dissimilar. The measured spectrum of Ni^{2+} is given in red.

5.6.4 Interpretation of the predictive part of the 3+2 OPLS model using coefficients rotated into pure spectrum profiles

The series of loadings spectra given in Figures 5.36 – 5.38 reveals that, while the OPLS model is unaffected by the presence of orthogonal variation, the PLS model captures this structure in terms of its third component. Now that we are aware that PLS is negatively affected by the orthogonal variation, we shall proceed by looking into other model parameters to uncover how sensitive they are to this contamination. We will start with the rotated coefficients and then later explore VIP.

As pointed out before, any OPLS model and its PLS counterpart may be expressed as a solution based on regression coefficients, and, in turn, these coefficients may be rotated such that they approximate the pure constituent profiles as well as possible. When the predictive regression coefficients (of the three predictive components) of the 3+2 OPLS model are rotated, we call the resulting vector *Y-related profiles* (plotted Figure 5.39). When the regression coefficients of the five components of the corresponding PLS model are rotated, we call the resulting vector *rotated coefficients* (plotted in Figure 5.40).

As is readily seen, the rotated coefficient vector of the PLS model (Figure 5.40) is contaminated by the Ni^{2+} peak, whereas the Y-related profile vector (Figure 5.39) of the 3+2 OPLS model is not. Hence, the breakdown of the rotated coefficients, and thereby the relevance of underlying PLS regression coefficients, in the presence of a large portion of orthogonal variation, is manifested. With a large fraction of orthogonal variation, only the OPLS method produces variable-related parameters that are fully consistent with reality.

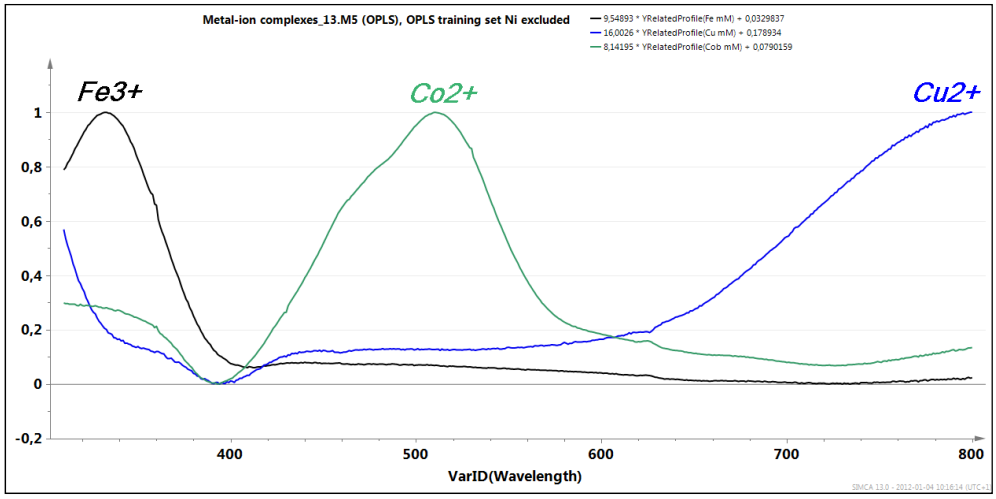


Figure 5.39: The pure spectral profiles of the three modeled metal-ion complexes. All data are scaled to fit the plot well.

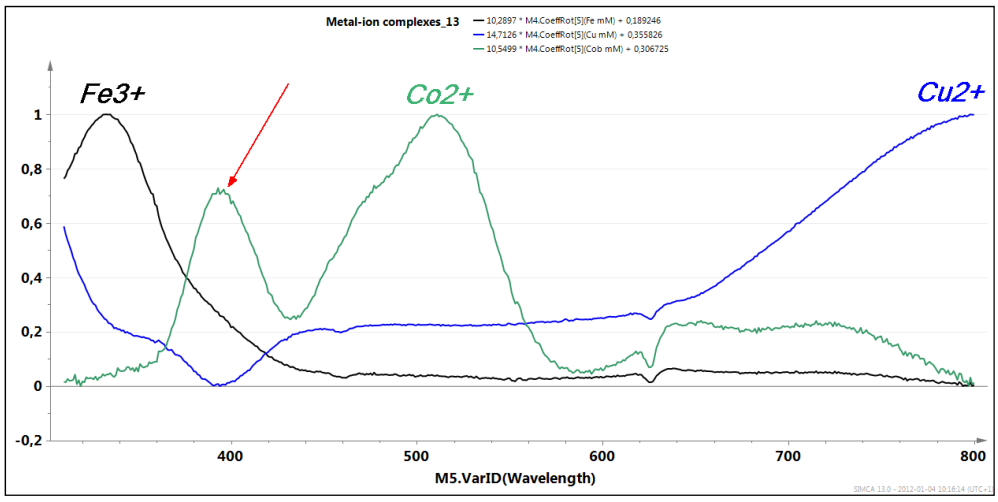


Figure 5.40: The rotated coefficients of the PLS model estimating the pure profiles of the three modeled metal-ion complexes. The ghost-peak, indicated by the arrow, in the region where Ni^{2+} absorbs light is caused by the orthogonal variation in the X-block. All data are scaled to fit the plot well.

5.6.5 Interpretation of the 3+2 OPLS model using VIP

In Section 5.5.8, Figure 5.33, two congruent VIP profiles were presented for the case when the Y-matrix consisted of four responses. This matching arose because the amount of orthogonal variation was small and therefore the OPLS and PLS models were very similar. Below, Figure 5.41 provides the VIP profiles for the current case with pronounced orthogonal variation. The VIP-values of the OPLS model and the PLS model are less similar than before, and particularly in the 360-440 nm region in which Ni^{2+} is absorbing.

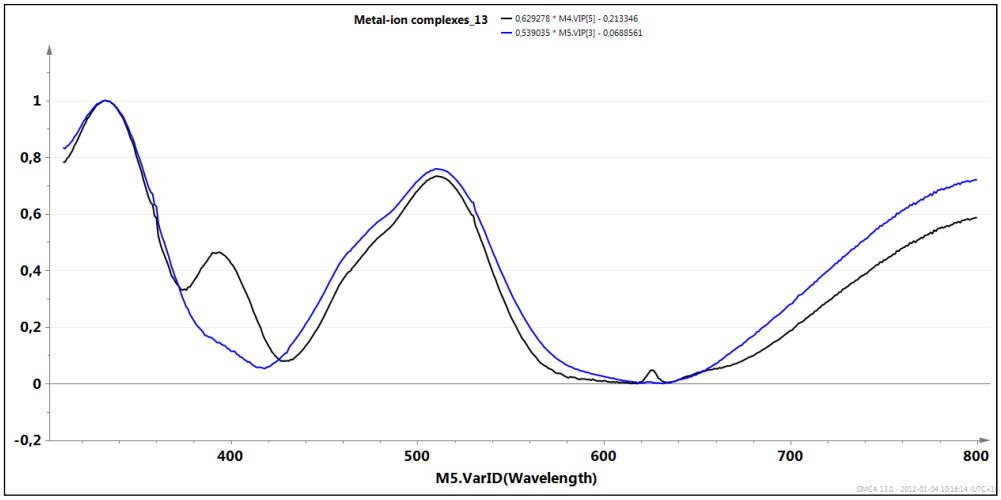


Figure 5.41: VIP profiles for the OPLS and PLS models fitted to three response variables. The VIP-values of the 3+2 OPLS model (blue) and the PLS model (black) are dissimilar, particularly in the 360-440 nm region. All data are scaled to fit the plot well.

5.6.4 Discussion of example

OPLS and PLS differ in the way that they handle orthogonal variation in X. PLS, on the one hand, incorporates such variation in the prediction model and this necessitates a rotation to ensure good predictions. Unfortunately, this compromises the ability of PLS to estimate pure spectral profiles. OPLS, on the other hand, avoids this problem by first removing the orthogonal variation from X before relating X to Y. The extent to which the estimates of the pure spectral profiles will differ between OPLS and PLS depends on how much orthogonal variation is present. However, OPLS will always yield similar or better pure profile estimates.

Interestingly, however, PLS and OPLS have maintained their predictive power after the exclusion of Ni^{2+} (no such results have been shown in Section 5.6). This underlines the ability of indirect calibration methods to handle unknown interferents in X. To ensure the best estimates of the pure profiles, such interfering species should be as uncorrelated with the other constituents as possible.

5.7 Questions for Chapter 5

1. What is the objective of OPLS?
2. What is meant by predictive component(s)? Orthogonal components?
3. Which variable-related parameter is useful to interpret a single-Y OPLS model?
4. Which variable-related parameters are redundant in a single-Y OPLS model?
5. Which variable-related parameters are useful to interpret a multi-Y OPLS model?
6. For what purpose can you use the Y-related profiles?

5.8 Summary and discussion

5.8.1 Some general remarks

OPLS represents a way to partition the variation in a data matrix X into systematic variation patterns unrelated to the response (orthogonal variation) and systematic variation patterns predictive of the response. Thus, OPLS enhances the interpretation of the PLS solution. All projection methods working in accordance with a least squares approach are sensitive to abnormalities in data. It is important to realize that detection and investigation of outliers represents an important part of multivariate data analysis and semi-empirical modeling. As discussed in Chapters 3 and 4, in both PLS and PCA, abnormal samples can be detected and

analyzed by looking at plots of scores, residuals and contributions. Outlier detection in OPLS is no different because exactly the same principles apply.

The first steps in OPLS are to estimate the loadings w and p which are obtained using the conventional partial least squares (PLS) method. The number of predictive and orthogonal components used should be selected according to some significance criterion such as cross-validation. It is important to use some significance criterion when determining the number of components. Underfitting or overfitting of data are serious problems with any empirical modeling technique and OPLS is no different. Remember that a single-Y PLS model, and its corresponding single-Y OPLS model with the same number of components, will have identical explained variances (R2). However, the predictive ability (Q2), as estimated by cross-validation, may differ slighly because the cross-validation routines used are different. Whereas (in SIMCA) PLS uses sequential cross-validation, OPLS uses full or total cross-validation (see also Section 4.7.3).

Furthermore, it is important to realize that the orthogonal variation in X removed from the calibration samples is assumed to exist in future samples too. This assumption stems from the fact that the updating parameters W_O^* and P_O are estimated from the calibration data only. No variation only present in future samples or calibration samples should be removed. This stresses the importance of removing only orthogonal components from the data.

5.8.2 The great asset of OPLS: Improved model transparency and interpretability

Perhaps the greatest advantage of OPLS is the improved knowledge and understanding that it provides. Imagine a twelve-component PLS model: Which variables relate to the response? Which plots and parameters are interesting? Some would analyze the regression coefficients, because they are used to predict Y from X. Others would use a combination of all the loadings, scores and coefficients, together with a check of the residuals. That sounds difficult but, given a good understanding of the data, usually works.

The OPLS method makes interpretation much easier. Firstly, it separates the predictive variation from the orthogonal variation. Secondly, it provides an opportunity to analyze the orthogonal variation and understand what might be causing it in the first place. Thirdly, the number of predictive PLS components is greatly reduced, making interpretation of the OPLS model much clearer.

Does this mean that conventional PLS models are uninterpretable without orthogonal filtering methods like OPLS? The answer is clearly no; for PLS models with few components (< 3), the overall interpretation is often very similar. The crucial property is the ratio of R2Xorth to R2Xpred (see Figure 5.1 for an illustration of the single-Y case). The larger this ratio, the more "disturbance" to the regular loading weights w_a^* and the more complicated the interpretation of a PLS model. This is because w_a^* expresses both the predictive and orthogonal variations in X. Model parameters derived from w_a^* like coefficients and VIP are affected in the same way.

Applying experimental design to both X and Y also facilitates interpretability. However, most PLS models containing more PLS components than the true rank of Y would benefit from using OPLS or any other orthogonal filtering method. Other inverse regression modeling methods e.g. PCR, MLR, and RR, are of course affected in the same way.

The suggested OPLS method is truly versatile, and, if properly used, will enhance data modeling and interpretation regardless of data type. Suppose that a dataset only contains relevant variation. Consequently, the OPLS method will not find any orthogonal components and the resulting OPLS model will converge to a conventional one component PLS solution. This is the case for data originating from designed experiments, where the columns are orthonormal to each other and no orthogonal latent variables are present. Also consider the opposite case; where the data consist only of orthogonal variation. Here, the OPLS method finds only orthogonal variation, no predictive components and therefore converges to the PCA solution.

5.8.3 Specific uses of OPLS

Instead of removing all systematic orthogonal variation in X, OPLS can be designed to remove specific types of variation in the X-data by setting the unwanted property as Y. A good example of this is temperature differences in NIR, producing undesirable spectral variation. OPLS allows the systematic spectral variation due to temperature to be analyzed, understood and removed.

OPLS can also be used to remove specific types of variation in process data acquired over time. This is achieved by setting e.g. a drift parameter measured in time as Y. Another example is QSAR, where thousands of molecular descriptors can be mechanically and effortlessly compiled. One problem with many of these descriptors is their dependence on molecular size. By setting a variable accounting for molecular size or volume as Y, OPLS can be used to filter out such dependencies from the suite of molecular descriptors. The size-corrected descriptors can thereafter be used to build QSAR models.

5.8.4 When and when not to use OPLS

Any data analysis, including PLS and OPLS, is based on an assumption of homogeneity and representativeness. Homogeneity means there are no outliers or strong sub-groups among the samples. Representativeness is achieved by selecting an informative and well-balanced training set. As outlined in Chapter 18, tools for DOE and multivariate design are invaluable in this respect. In cases where these important assumptions are violated, there is a substantial risk that PLS, OPLS, and similar models, become overfitted. In such cases, the models should only be used with caution.

Thus, in an OPLS model it is important to scrutinize the predictive and orthogonal components carefully to reveal whether the assumptions of homogeneity and representativeness are indeed fulfilled. Are there any groups, outliers, non-linearities or time trends? Such features are undesirable as their presence may lead to an overfit and reduced model interpretability. Strong groupings in the Y-orthogonal part also questions the relevance of deriving just one overall calibration model.

A strong indication of the suitability of OPLS is when the amount of variance explained in Y is relatively small on the first component whilst the proportion of variance explained in X is rather large. This suggests that significant X variation may exist that is orthogonal to Y.

5.8.5 Related methods and some extensions

A number of different algorithms for orthogonal modeling exist, for a compilation see e.g. [Svensson, *et al.*, 2002]. An early precursor method, called Target Projection, was developed by Kvalheim in the mid 80s [Kvalheim, 1989]. From this work it was noticed that PLS and other latent variable regression methods could be reduced to a single predictive latent variable by using the regression vector as loading weight. More recently, Josefson and coworkers proposed an elegant modification denoted Transposed-OPLS [Abbas, *et al.*, 2011]. Transposed-OPLS is intended to enhance the analysis and interpretation in applications handling hyperspectral-imaging data. Using this procedure, spectra are transposed such that every spectrum becomes a column in the data table instead of a row. Thereafter OPLS is applied between the matrix with the transposed spectra and the reference spectrum, the latter of which is used as the single Y-variable. The reference spectrum relates to a target compound, biomolecule, or metabolite occurring in the cells of a living system, and for which it is of interest to accomplish an understanding of its distribution inside the living system. Hence, the predictive component will capture everything in the measured spectra (making up the hyperspectral image) that correlates with the reference spectrum. This makes it possible to visualize the distribution of the compound coupled to the reference spectrum inside the living system.

Moreover, Trygg has also modified the two-block X/Y OPLS approach to improve its interpretational functionality [Trygg, 2002; Trygg and Wold 2003]. This approach is called O2PLS. In contrast to OPLS, which is uni-directional ($X \rightarrow Y$), O2PLS is bi-directional in the sense that it is predictive both ways, i.e. $X \leftrightarrow Y$. This implies that with O2PLS one can

model (i) the joint X-Y covariation, (ii) the Y-orthogonal variation in X, *and* (iii) the X-orthogonal variation in Y. O2PLS and some extensions are discussed in Chapter 6.

5.8.6. Concluding remarks

The case studies reported in this chapter suggest that OPLS represents the best regression method for the future, as it simultaneously offers good predictive ability and enhanced interpretability of complicated multivariate systems. In the case of a single-Y problem, OPLS assembles the predictive X-variation in the first component, and, when dealing with multi-Y problems, in the first few predictive components. When faced with very wide datasets, involving thousands of variables, the extension of OPLS to hierarchical-OPLS is an interesting prospect. One may envisage separate hierarchical models for the predictive and orthogonal parts of the X-variation.

OPLS is capable of estimating pure profiles of constituents in a mixture as well as direct calibration methods. Interestingly, this ability to estimate pure spectra is not restricted to multivariate calibration problems, but applies in any situation in which the Y-variables are known to, or can be assumed to, contribute to X in an additive manner.

6 O2PLS

6.1 Objective

In this chapter we discuss a second modification to the standard PLS method, denoted O2PLS. The primary objective of O2PLS is *integration* of data in the two-block (X/Y) context and understanding of which information overlaps between the two data tables and which information is unique to a specific data table (X or Y). O2PLS accomplishes this task by a flexible model structure incorporating three types of components, that is, (i) components expressing the joint X/Y information overlap, (ii) components expressing what is unique to X, and (iii) components expressing what is unique to Y. This means that when giving reference to an O2PLS model, it becomes necessary to quote a model consisting of A+B+C components, all corresponding to the different compartments of the model. The result is a model with improved interpretability. In order to illustrate O2PLS, we will use two example datasets. The first illustration is a chemical process dataset where the information in spectral measurements and classical process variables is analyzed. The second illustration is a multivariate calibration example involving multiple blocks of spectral data.

6.2 Introduction

Chapters 4 and 5 have outlined projection methods targeting the two-block (X/Y) *regression* problem, i.e., PLS and OPLS. Using the information in X, these methods construct a predictive model of Y ($X \rightarrow Y$), which can predict new Ys for modified Xs. By way of examples in Chapter 5, we have demonstrated that these methods provide very similar models as long as the orthogonal variation in X is small. With increasing proportions of orthogonal variation in X, however, the outcomes of the two methods start diverging, and eventually OPLS will produce models that are more realistic to interpret. The score-loading correspondence in PLS is perturbed by the presence of larger proportions of orthogonal variation in X.

Interestingly, in conjunction with the OPLS approach also comes another extension to the conventional PLS method, denoted O2PLS [Trygg, 2002; Trygg and Wold, 2003]. Unlike PLS and OPLS, however, O2PLS does *not* target the regression problem; instead, it is more suited to problems where the goal is to accomplish an overview and to integrate the information in multi-block data structures. The most parsimonious models arise for the two-block problem, but the methodology can be extended in various ways to address very complex datasets, possibly organized as multiple blocks or databases. We shall here illustrate O2PLS with two datasets, the first consisting of two blocks of data and the second comprising multiple blocks of spectral data. As will be shown via these examples, the scope of O2PLS is more in line with the scope of PCA, the latter of which strives to overview and summarize the information in measured data. Recall that PCA was introduced in Chapter 3.

Considering the two-block (X/Y) situation, we can state that O2PLS is a bidirectional method, i.e. $X \leftrightarrow Y$, and the "2" in the name of the method derives from this bidirectionality property. Conceptually, this means that X can be used to model Y, *and* that Y can be used to model X. Practically, the net result is that O2PLS allows the partitioning of the systematic variability in X and Y into three parts; (i) the information that is simultaneously present in X and Y, (ii) the information that is only found in X, and (iii) the information that is only found Y. Another consequence of the O2PLS model setup is that it

becomes necessary to refer to a model consisting of A+B+C components, all corresponding to the different compartments of the model. It should be noted that in some cases the B components and the C components are estimated by different methods, making the model consisting of A+B1+B2+C1+C2 components. This notation is further elaborated in Section 6.3.3.

Moreover, the switch in operational focus between OPLS (regression between two blocks of data) and O2PLS (integration of two blocks of data) is accompanied by a transition in the preferred terminology. Whereas with OPLS it is practical to distinguish between *predictive* and *orthogonal* sources of variability, when using O2PLS it becomes natural to consider *joint* and *unique* information structures. The joint variation represents the information overlap between X and Y, i.e., the systematic structure that is present in both X and Y. Any unique variation corresponds to information beyond the overlap and which is exclusively found in either the X- or the Y-block. An overview of how O2PLS divides X- and Y-variabilities is displayed in Figure 6.1.

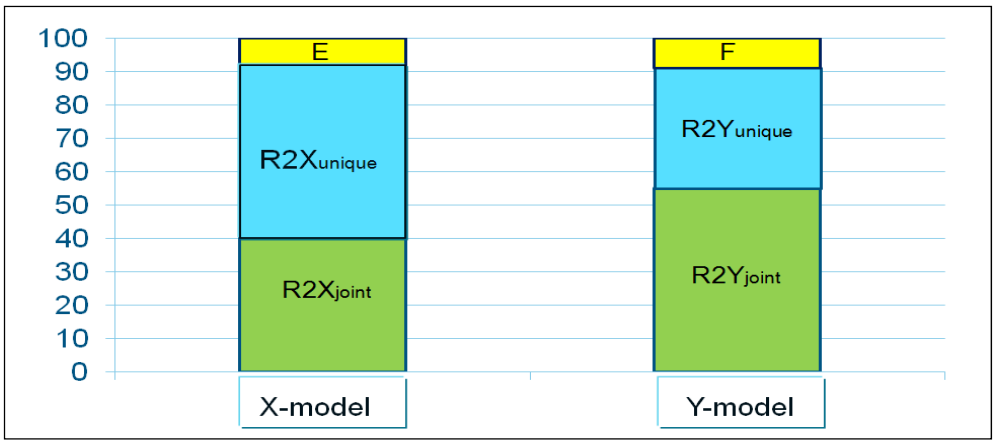


Figure 6.1: O2PLS finds the joint variation in X ($R2X_{joint}$) and Y ($R2Y_{joint}$), which represents the information overlap between X and Y. Furthermore, O2PLS finds the unique variation in X ($R2X_{unique}$) and the unique variation in Y ($R2Y_{unique}$), which correspond to information that is exclusively found in either X or Y. Finally, the X- and Y-variabilities that O2PLS cannot cope with go to the E and F residual structures.

There is no doubt there is a rising interest for the O2PLS methodology in research, development and production. The technique has found important use in such diverse fields as multivariate calibration, process monitoring, quantitative structure-activity relationship modeling, plant metabolomics and clinical applications [see, e.g., Trygg and Wold, 2003; Gabrielsson, *et al.*, 2006; Bylesjö, *et al.*, 2007; Stenlund, *et al.*, 2009; Suzuki, *et al.*, 2011; Eriksson, *et al.*, 2012; Frost, *et al.*, 2013]. In the following, the O2PLS method will be explained and exemplified. First, we shall discuss some theory of O2PLS. Then, the O2PLS method is illustrated by means of two examples.

6.3 Theory of O2PLS

6.3.1 Objective and structure of O2PLS

As indicated by Figure 6.1, the objective of O2PLS is to discover *joint* and *unique* variabilities in X and Y. This is accomplished by a model structure made up from as many as five types of components. Figure 6.2 outlines the core matrix structures of the O2PLS model and the related equations are given in Section 6.3.2. Here, X and Y are the two data tables to be integrated. Their joint variation is expressed by the X-scores T, the Y-scores U, the X-loadings P' and the Y-loadings Q'. In the SIMCA software, these parameters are retrievable from the so called *predictive* components. Moreover, the variation that is unique to X is stored in the scores T_O and the loadings P_O'. The subscript *O* is utilized to indicate that these parameters are given by the *orthogonal-in-X* components in SIMCA. Similarly, the scores U_O and the loadings Q_O' contain the variation that is unique to Y, and the subscript *O* designates the *orthogonal-in-Y* components in SIMCA. It should be observed

that all the orthogonal components are estimated as part of the O2PLS algorithm or by means of a final PCA step. This is further elaborated below. Finally, we have the unsystematic part of the data, the residuals, which we call the E and F residual matrices. Parameters like DModX and DModY are derived from E and F, see more information on these in Section 4.7.1.

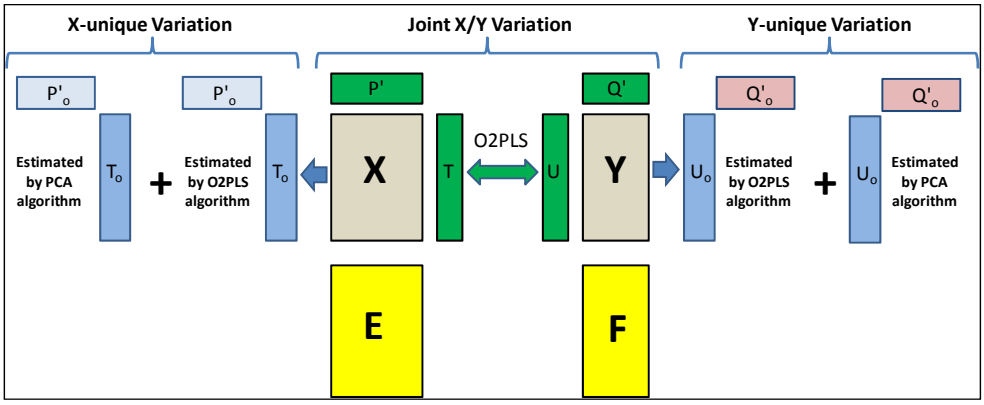


Figure 6.2: An overview of the O2PLS model. The O2PLS model has components that express (i) the joint X/Y information overlap, (ii) the variation that is unique to X, and (iii) the information that is unique to Y. In addition, there are the two residual matrices, denoted E and F, dedicated to the unexplained X- and Y-variabilities, respectively. See text for more details.

6.3.2 The O2PLS model

Referring to Figure 6.2, the O2PLS model can be written as follows:

$$\mathbf{X} = \mathbf{1}\bar{\mathbf{x}}' + \mathbf{TP}' + \mathbf{T_oP_o}' + \mathbf{E} \qquad \text{(eqn. 6.1)}$$

$$\mathbf{Y} = \mathbf{1}\bar{\mathbf{y}}' + \mathbf{UQ}' + \mathbf{U_oQ_o}' + \mathbf{F} \qquad \text{(eqn. 6.2)}$$

It is the matrix products **TP'** and **UQ'** that hold the joint X/Y information overlap. The number of score vectors (in **T** and **U**) and loading vectors (in **P'** and **Q'**) is determined using cross-validation. In the example Model Window below (Figure 6.3) there are six components of this type (predictive components).

Furthermore, the unique variations in X and Y are accounted for by the matrix products $\mathbf{T_oP_o}'$ and $\mathbf{U_oQ_o}'$, respectively. It should be noted that the components belonging to these model parts are estimated through the O2PLS algorithm and/or the PCA algorithm in SIMCA. Thus, $\mathbf{T_oP_o}'$ and $\mathbf{U_oQ_o}'$ may be decomposed as:

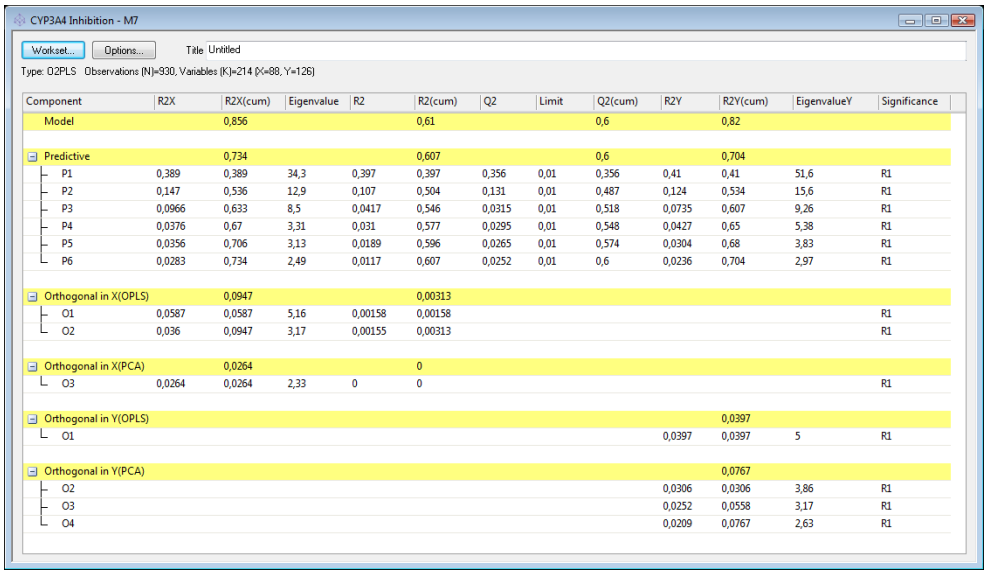
$$\mathbf{T_oP_o}' = \mathbf{T_{o\text{-}O2PLS}P_o}'_{\text{-}O2PLS} + \mathbf{T_{o\text{-}PCA}P_o}'_{\text{-}PCA} \qquad \text{(eqn. 6.3)}$$

$$\mathbf{U_oQ_o}' = \mathbf{U_{o\text{-}O2PLS}Q_o}'_{\text{-}O2PLS} + \mathbf{U_{o\text{-}PCA}Q_o}'_{\text{-}PCA} \qquad \text{(eqn. 6.4)}$$

Subsequent to the convergence of the O2PLS algorithm, PCA is independently applied to the existing X- and Y-residuals to make sure they are fully exhausted, i.e., that all systematic structures are captured and only non-systematic residuals remain at the end. These components are listed in the Model Window (see below in Figure 6.3). The example lists 3 (2+1) orthogonal-in-X and 4 (1+3) orthogonal-in-Y components and their origin is indicated. Irrespective of derivation principle (algorithm), the number of components is estimated through cross-validation.

6.3.3 Implementation in SIMCA

For each computed O2PLS model, the performance statistics are summarized by means of a Model Window. An example Model Window is shown in Figure 6.3. All components come out in descending order of importance, i.e., they explain lesser and lesser of the variances.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,856			0,61			0,6		0,82		
Predictive		0,734			0,607			0,6		0,704		
P1	0,389	0,389	34,3	0,397	0,397	0,356	0,01	0,356	0,41	0,41	51,6	R1
P2	0,147	0,536	12,9	0,107	0,504	0,131	0,01	0,487	0,124	0,534	15,6	R1
P3	0,0966	0,633	8,5	0,0417	0,546	0,0315	0,01	0,518	0,0735	0,607	9,26	R1
P4	0,0376	0,67	3,31	0,031	0,577	0,0295	0,01	0,548	0,0427	0,65	5,38	R1
P5	0,0356	0,706	3,13	0,0189	0,596	0,0265	0,01	0,574	0,0304	0,68	3,83	R1
P6	0,0283	0,734	2,49	0,0117	0,607	0,0252	0,01	0,6	0,0236	0,704	2,97	R1
Orthogonal in X(OPLS)		0,0947			0,00313							
O1	0,0587	0,0587	5,16	0,00158	0,00158							R1
O2	0,036	0,0947	3,17	0,00155	0,00313							R1
Orthogonal in X(PCA)		0,0264			0							
O3	0,0264	0,0264	2,33	0	0							R1
Orthogonal in Y(OPLS)										0,0397		
O1									0,0397	0,0397	5	R1
Orthogonal in Y(PCA)										0,0767		
O2									0,0306	0,0306	3,86	R1
O3									0,0252	0,0558	3,17	R1
O4									0,0209	0,0767	2,63	R1

Figure 6.3: Model Overview Window for a 6+2+1+1+3 O2PLS model. For explanation, please see the text.

The Model Window in Figure 6.3 relates to a 6+2+1+1+3 O2PLS model. The example model consists of 6 *predictive* components summarizing the joint X/Y variation, 3 (2+1) *orthogonal-in-X* components holding what is unique to X, and 4 (1+3) *orthogonal-in-Y* components holding what is unique to Y. Recall from the foregoing section that the orthogonal components can be estimated by both the O2PLS algorithm and the PCA algorithm. More specifically, the Model Window in Figure 6.3 includes:

- a top row summarizing the model, showing the cumulative R2X, R2, Q2, and R2Y;
- predictive components: one row per significant component, showing the joint X/Y variation;
- orthogonal components in X: one row per significant component, showing the variation in X that is unique to X;
- orthogonal components in Y: one row per significant component, showing the variation in Y that is unique to Y.

The table columns are:

- Component - Component index.
- R2X - Fraction of X variation modeled in the component, using the X model.
- R2X(cum) - Cumulative R2X up to the specified component.
- Eigenvalue - The minimum number of observations (N) and X-variables (K) multiplied by R2X.
- R2 - Fraction of Y variation modeled in the component, using the X model.
- R2(cum) - Cumulative R2 up to the specified component.
- Q2 - Fraction of Y variation predicted according to cross-validation in the component, using the X model.
- Q2(cum) - Cumulative Q2 up to the specified component.
- R2Y - Fraction of the Y variation modeled in the component, using the Y model.
- R2Y(cum) - Cumulative R2Y up to the specified component.

- Eigenvalue Y - The minimum number of observations (N) and Y-variables (M) multiplied by R2Y.
- Significance - CV significant or insignificant according to rule

Given its objective of data integration, in comparison with the standard two-block (X/Y) PLS method, the O2PLS model yields far more model performance indicators (R2/Q2…) and model parameters (scores/loadings…). Most likely, it will take some time for the proponents of PLS to realize the potential of O2PLS and adapt their modeling accordingly. In order to catalyze this transition, SIMCA provides a newly developed plot denoted OPLS/O2PLS Overview, which provides a graphical summary of the joint and unique variabilities. Being an extension of the conventional PCA/PLS Summary of Fit plot of SIMCA, this column plot is both stacked and color coded in accordance with the magnitudes of the joint and unique variance structures (Figure 6.4). One can say that Figure 6.1 is a "light" version of Figure 6.4.

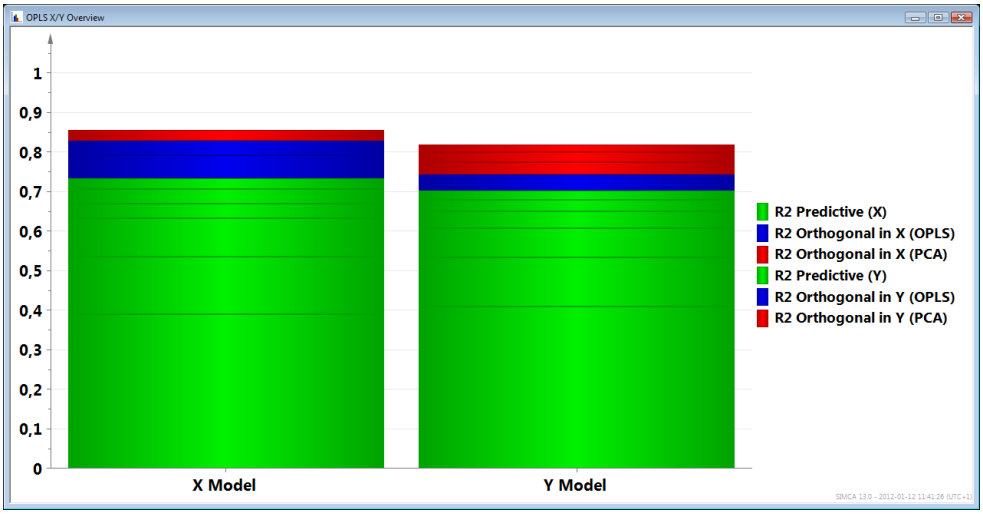


Figure 6.4: The OPLS/O2PLS overview plot. Conceptually, this plot is a more elaborate version of the graph shown in Figure 6.1. Thus, this column plot is stacked and color coded to indicate the magnitudes of the joint and unique variabilities in X and Y. The green color corresponds to the joint X/Y variation. The blue color indicates unique variations in X or Y, estimated in terms of the OPLS/O2PLS algorithm. The red color represents unique variations in X or Y, estimated by PCA subsequent to the convergence of the OPLS/O2PLS algorithm.

With the theory of O2PLS and the implementation of the method in the SIMCA software in mind, we now turn to the two illustrations, described in Sections 6.4 and 6.5.

6.4 Integrating spectral and process data of a batch process

With the increasing availability of different process analysers and sensors, multiple data sources are commonly available and this will impose new challenges and enable new types of information-extraction possibilities in data mining of complex process data. The ability to separate joint and unique, and maybe even redundant, information in data organized as multiple blocks or data bases will be of increasing importance. This enables a detailed assessment of both the joint and the unique variations in the respective datasets, and eventually a deeper understanding of the entire process or system. Our first example has been chosen to illustrate the analytical potential of O2PLS in comparing and integrating process data originating from different sources.

6.4.1 The batch hydrogenation dataset

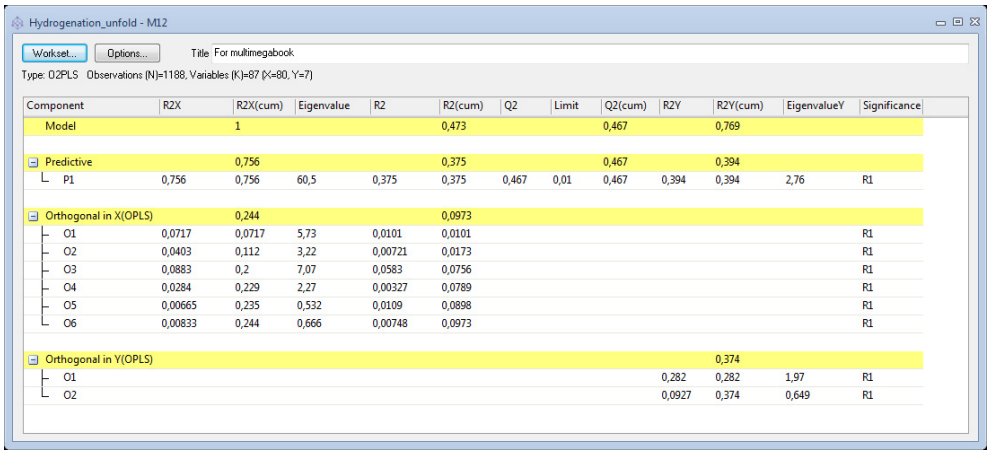
The example relates to a chemical reaction run as a batch process [Gabrielsson, *et al.*, 2006]. The reaction is a hydrogenation, triggered using H_2 gas flow, converting nitrobenzene to aniline. Six reaction batches, here denoted CP1-CP6, were run under identical conditions as part of a multivariate statistical process control (MSPC) study. The

time for the reaction to run to completion was around 35 minutes. Process and spectral data were recorded every 10 seconds, giving a total of 1188 observations upon unfolding the three-way batch data table to a two-way data table preserving the direction of the variables. More information on this unfolding principle can be found in the chapter on batch statistical process control, BSPC (Chapter 16).

Seven process variables were measured: reactor temperature, reactor pressure, gas feed, jacket-in temperature, jacket-out temperature, flow rate of oil, and stirrer speed. UV-spectra were recorded in the interval 220-290 nm and each spectrum was pre-processed using the first derivative. This resulted in 80 spectral variables. Prior to the O2PLS modeling, the spectral data (here used as the X-block) were mean-centered and the process data (here used as the Y-block) were mean-centered and column-scaled to unit variance.

6.4.2 O2PLS analysis of the spectral and process data

The O2PLS analysis of the batch hydrogenation dataset gave a model with nine (1 + 6 + 2) components, divided as one predictive component, six orthogonal-in-X components and two orthogonal-in-Y components. The Model Window is seen in Figure 6.5. As shown by Figure 6.5, all the orthogonal components are in this case estimated by the O2PLS algorithm; no additional orthogonal components were found when applying PCA to the E and F residual matrices. Inspection of Figure 6.5 shows that the spectral data (X) contain more joint information than the process data (Y), i.e., 75.6% of the variance in the spectral data overlap with 39.4% of the variance in the process data. This is according to expectation as the spectral data closely monitors the progress of the chemical reaction. The process data contain more noise and the O2PLS model leaves as much as 23.1% (100 - 76.9) of the Y-variance in the residuals. This high number may be due to, for example, flow rate of oil for cooling which is regulated at short irregular intervals. Orthogonal variation is likely more common in process data than spectroscopic data, since the former may contain manipulated variables, for example, temperature control. Controlled variables often include structured trajectories which are not necessarily fully correlated to the progress of the reaction or the process.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		1			0,473			0,467		0,769		
Predictive		0,756			0,375			0,467		0,394		
P1	0,756	0,756	60,5	0,375	0,375	0,467	0,01	0,467	0,394	0,394	2,76	R1
Orthogonal in X(OPLS)		0,244			0,0973							
O1	0,0717	0,0717	5,73	0,0101	0,0101							R1
O2	0,0403	0,112	3,22	0,00721	0,0173							R1
O3	0,0883	0,2	7,07	0,0583	0,0756							R1
O4	0,0284	0,229	2,27	0,00327	0,0789							R1
O5	0,00665	0,235	0,532	0,0109	0,0898							R1
O6	0,00833	0,244	0,666	0,00748	0,0973							R1
Orthogonal in Y(OPLS)										0,374		
O1									0,282	0,282	1,97	R1
O2									0,0927	0,374	0,649	R1

Figure 6.5: Model window of the 1 + 6 + 2 O2PLS model for the batch hydrogenation dataset.

6.4.3 Interpretation of the joint variation

The scores and loadings of the single predictive component are plotted in Figures 6.6 – 6.9. Figure 6.6 provides the line plot of the X-score t for the joint variation in the spectroscopic data when related to the process data. Overall the score trajectories are congruent between the six batches. There is an initialization phase for the conversion reaction lasting about 10% into the life-time of the batch. Subsequently, there is a smooth linear development in the score trajectory corresponding to gradually decreasing contents of nitrobenzene (reactant) and increasing proportions of aniline (product). The corresponding X-loadings p are plotted in Figure 6.7. Recall that the raw spectra were pre-processed by taking the first derivative, so the loading profile is expressed in this metric. A comparison of the derivated

spectra (no plot shown) and the loading profile of Figure 6.7 confirms the interpretation of the score vector t as reflecting changing concentrations of the starting compound and the resulting product.

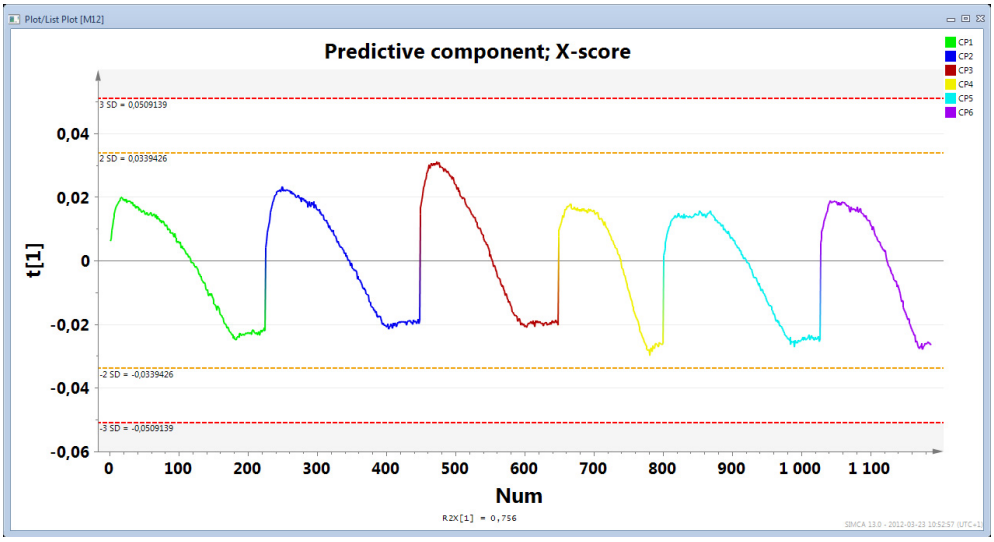


Figure 6.6: Line plot of the X-score t for the single predictive component. The plot is color coded according to batch.

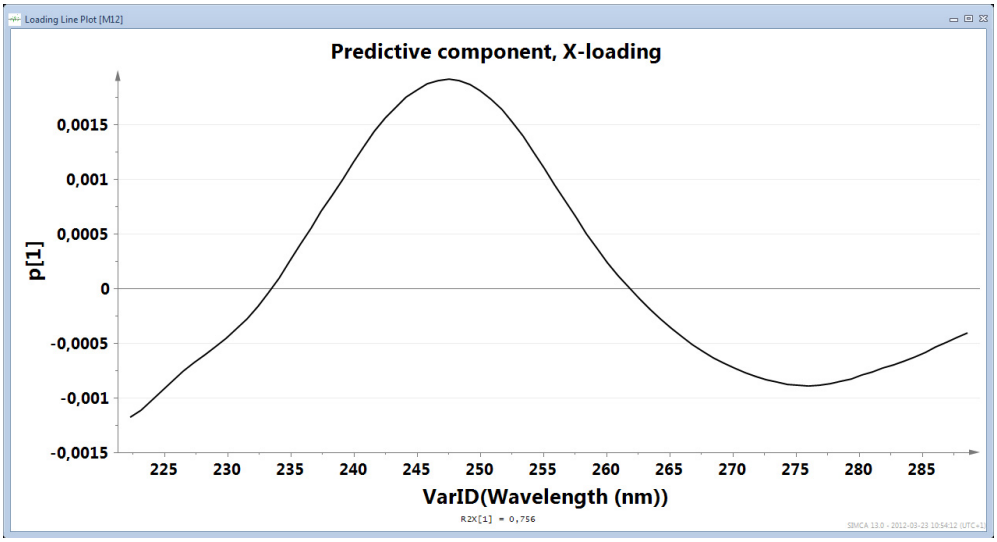


Figure 6.7: Line plot of the X-loading p for the single predictive component.

Figures 6.8 and 6.9 show the scores and loadings for the joint variation in the process data when related to the spectroscopic data. The scores in Figure 6.8 do not have the same smoothness as observed in Figure 6.6. This is due to sudden changes in some of the trajectories of the process variables. Figure 6.9 indicates that gas feed, reactor temperature, and temperatures for jacket in and jacket out predominantly have a linear association with the spectroscopic data. The model interpretation shows that gas feed is decreasing towards the completion of the reaction. The jacket temperatures are controlled to maintain a pre-defined reaction temperature profile and both variables increase numerically with the progress of the reaction.

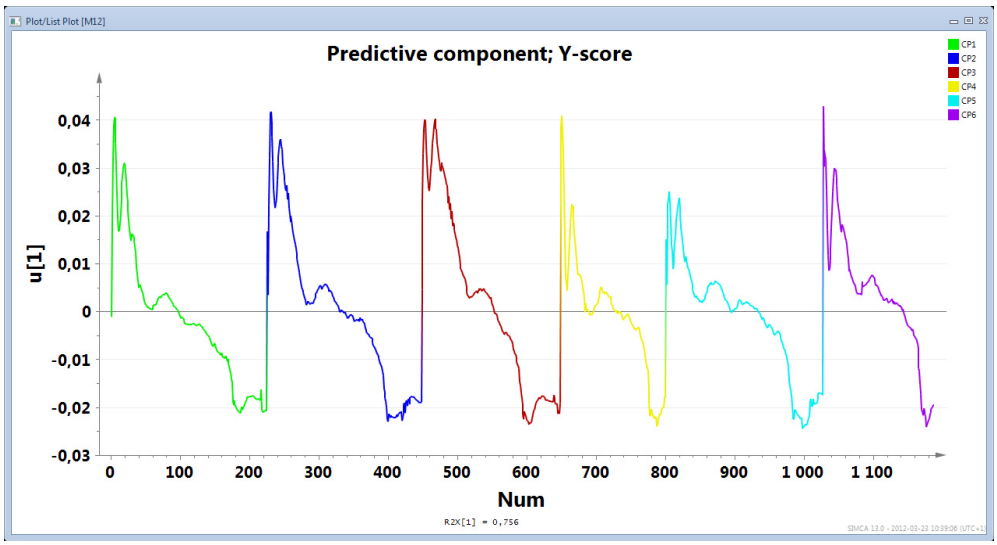


Figure 6.8: Line plot of the Y-score u for the single predictive component. The plot is color coded according to batch.

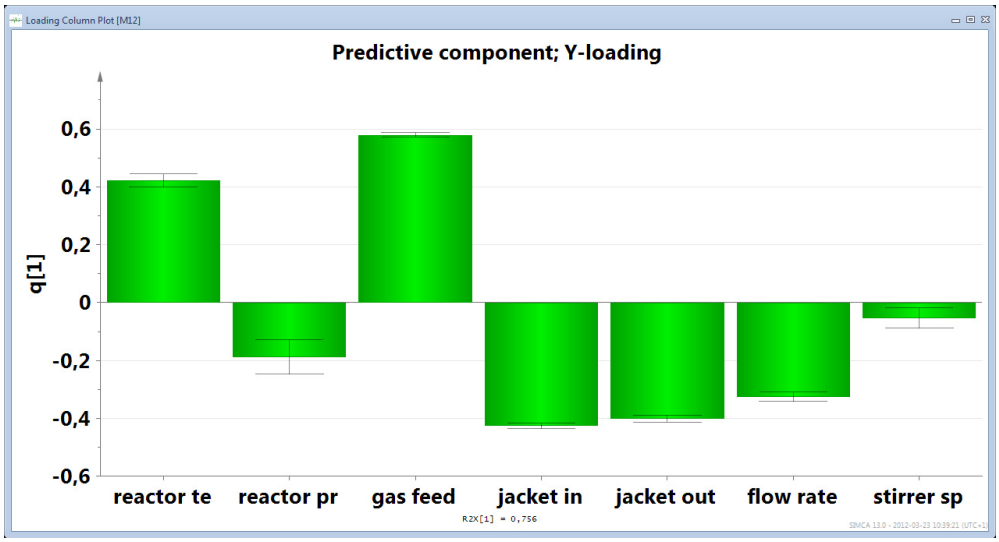


Figure 6.9: Column plot of the Y-loading q for the single predictive component. The error bars correspond to approximate 95% confidence intervals.

6.4.4 Interpretation of the unique variation in the spectral data

The systematic variation that is only found in the spectral data is described by the orthogonal-in-X components. As shown by Figure 6.5, there are six components of this type that jointly model 24.4% of the spectral variance. We shall here focus the interpretation on the first of these. Figure 6.10 is a line plot of the score vector of the first orthogonal-in-X component. Evidently, batches CP2 and CP5 deviate in the score plot, batch CP2 with consistently more positive score values than the other batches, and batch CP5 with always more negative scores. The corresponding loading profile (Figure 6.11) is similar to the loading profile of the joint variation (Figure 6.7). This similarity implies that the first orthogonal-in-X component captures structures between the spectral and process data that are non-linear in nature. The between batch variability modeled by this component is due to differences in reaction rate and amount of product formed. Batch CP5 has a trajectory similar to the other batches (CP1, CP3, CP4 & CP6), but the reaction yields less product. Batch CP2 has a different trajectory suggesting faster reaction progression and higher amount of product.

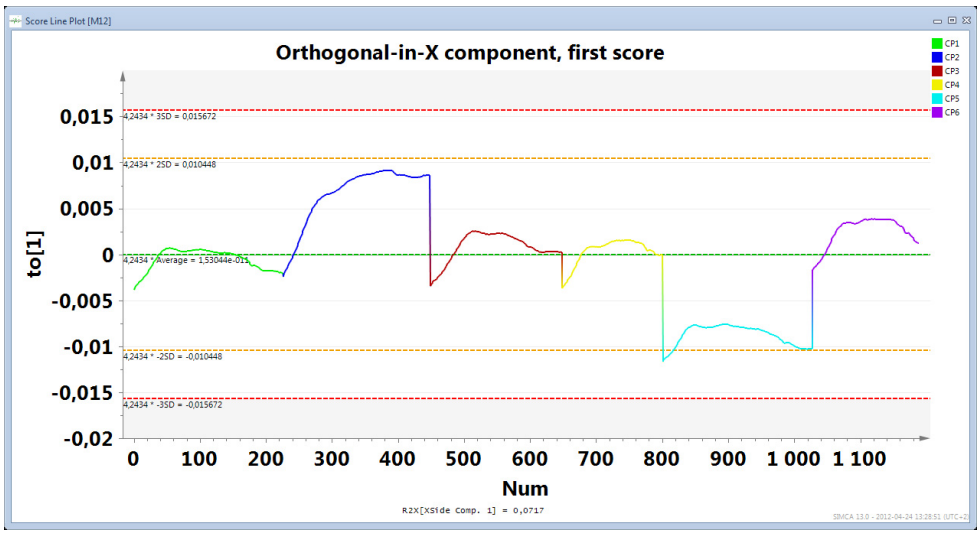


Figure 6.10: Line plot of the X-score to for the first orthogonal-in-X component. The plot is color coded according to batch.

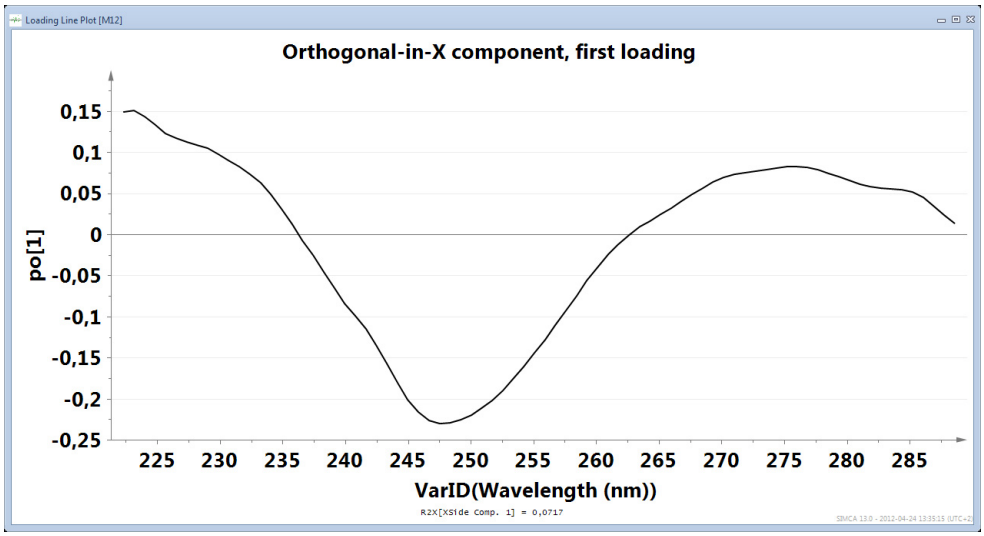


Figure 6.11: Line plot of the X-loading po for the first orthogonal-in-X component.

6.4.5 Interpretation of the unique variation in the process data

The systematic variation that is only found in the process data is described by the two orthogonal-in-Y components, which account for 37.4% of the variance in the seven process variables. Figure 6.12 provides a line plot of the score vector of the first orthogonal-in-Y component. This component is strongly influenced by early fluctuations in the reactor pressure. The influence of the reactor pressure can be deduced from the loadings plotted in Figure 6.13.

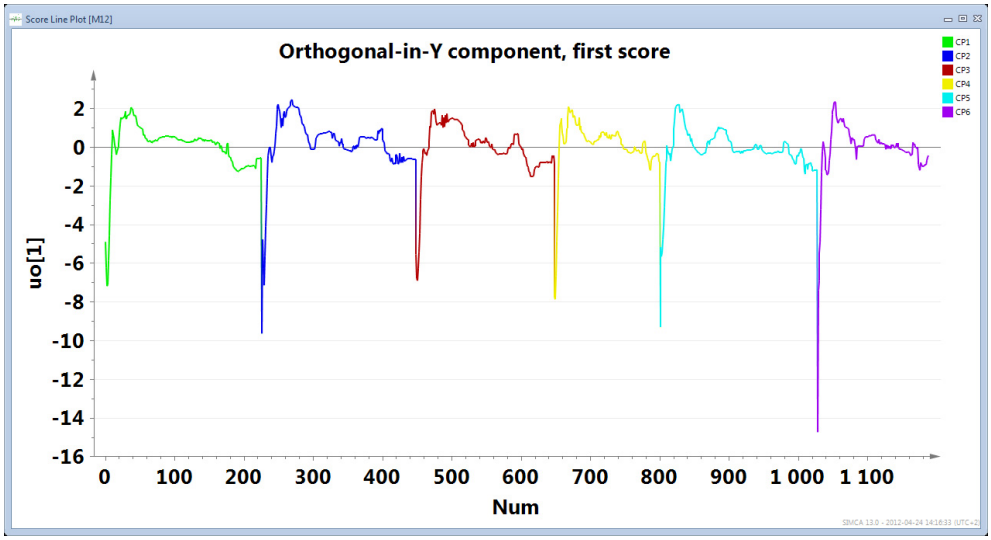


Figure 6.12: Line plot of the Y-score uo for the first orthogonal-in-Y component. The plot is color coded according to batch.

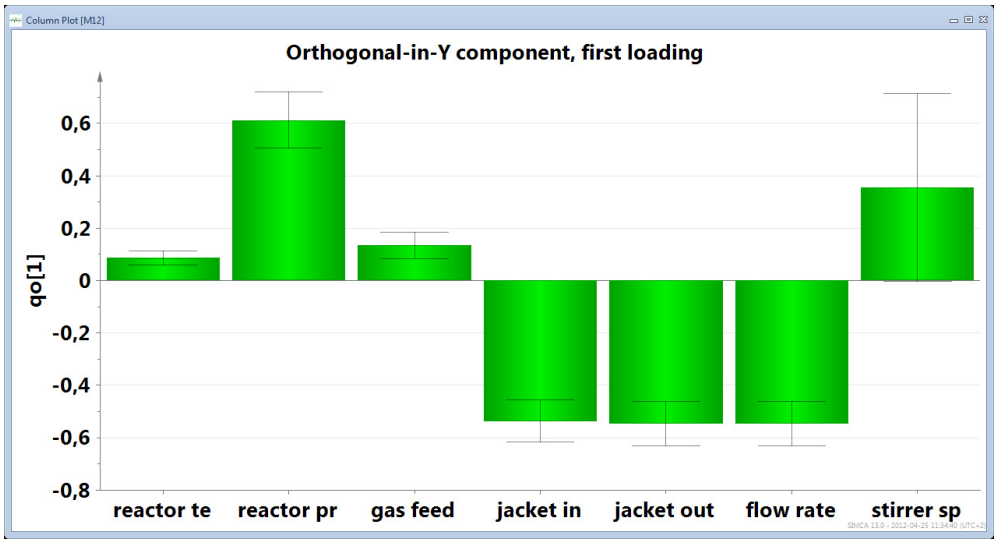


Figure 6.13: Line plot of the Y-loading qo for the first orthogonal-in-Y component.

6.4.6 Summary of the batch hydrogenation dataset

The O2PLS model shows that 75.6% of the spectral variation and 39.4% of the variation in the seven process variables overlap. The spectral data describes the evolution of the hydrogenation reaction and hence it is not surprising that a large part of the spectral variation is accounted for by the predictive component. The variation that is unique to the spectral measurements reflect small between-batch differences due to variations in the kinetics of the chemical reaction. This conclusion is based on the fact that the joint and unique loading profiles (compare Figures 6.7 and 6.11), albeit with an inverse relationship, are similar. In case they are not, the unique variations could indicate e.g. a competing side reaction.

It is interesting to compare the predictive Y-loadings plotted in Figure 6.9 and the orthogonal-in-Y loadings plotted in Figure 6.13. The seven process variables can be divided in three categories. The first category is made up of reactor temperature and gas feed, which have a strong influence on the evolution of the chemical reaction. Hence, they have a strong linear relationship to the spectral data, and therefore mainly contribute to the single predictive component of the O2PLS model. The second group of process variables consists of reactor pressure and stirrer speed, both of which express variation that is unique and not

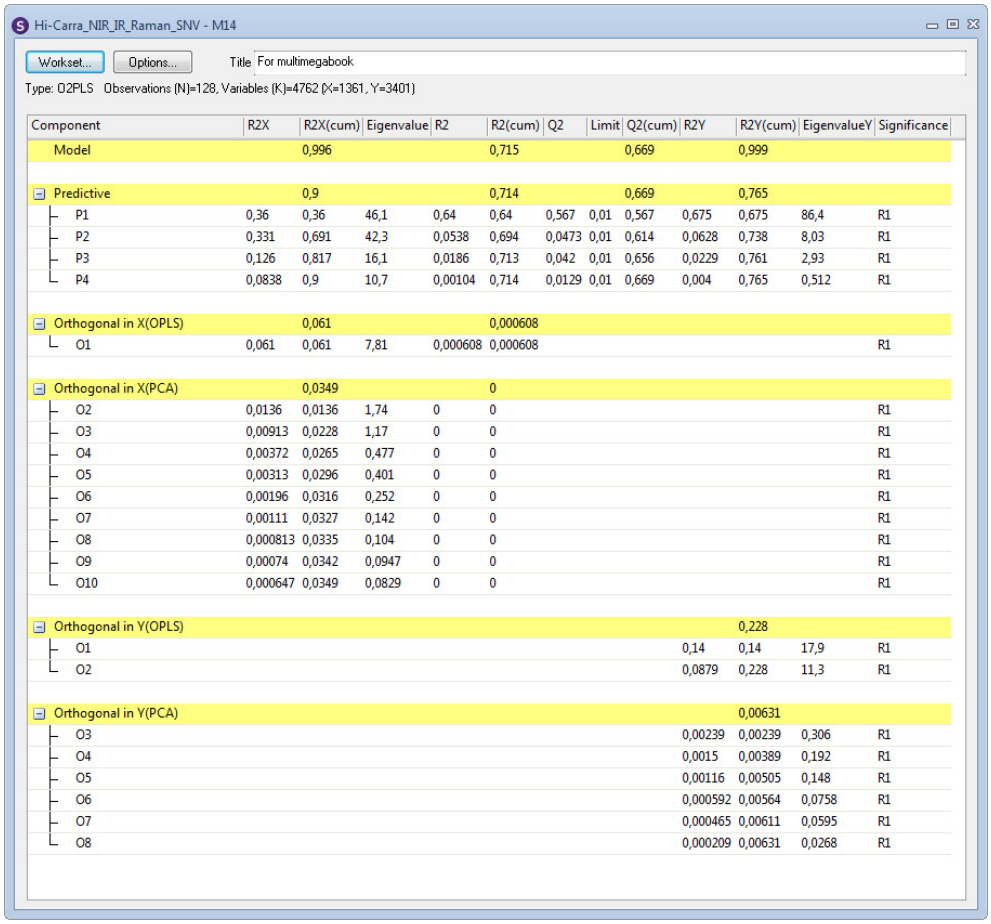
contained in the spectral measurements. Much of the variability in these two process variables are unrelated to the general progress of the hydrogenation reaction. The third group comprises flow rate of oil and the jacket in and jacket out temperatures. The latter variables have an effect on the reaction and the amount of product formed, but at the same time they contain systematic variability not linearly related to the spectral data. Variability in process variables that is unique and not found in the spectral data could e.g. arise in the case of controlled variables for which settings are attempted that should minimize some undesired effect of the reaction, for example, evolution of heat. The short and sometimes irregular behavior associated with such process variables is not correlated to the general development of the chemical reaction.

In summary, the O2PLS analysis of the spectral and process data offers a good summary of the process behavior and a deeper understanding of what are the relationships and what are not the relationships between the two sets of data. The grouping of the seven process variables into three subsets was possible through the interpretation of the joint and unique variabilities. Detecting similar subset formation would have been much more difficult using PLS, since PLS lacks the ability to differentiate between joint and unique information structures.

6.5 Integrating multiple spectral datasets in a multivariate calibration study

The second illustration is an example of multivariate calibration based on multiple spectral datasets. For a series of 128 observations (samples) spectral measurements were made using NIR, IR and Raman instruments. It was of interest to explore the information overlap between the analytical platforms. A more thorough introduction to this dataset is provided in the chapter on hierarchical methods (Chapter 19). We are indebted to Jan Larsen, CPKelco A/S, Denmark, and co-workers for granting us permission to use this dataset. Original reference: M. Dyrby et al. Carbohydrate Polymers 57, 337-348, 2004.

In the O2PLS analysis, the NIR and IR data were merged to form one block of data (X, 1361 variables). The Raman data were used as the second block of data (Y, 3401 variables). Figure 6.14 shows the model window of the resulting O2PLS model.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,996			0,715			0,669		0,999		
Predictive		0,9			0,714			0,669		0,765		
P1	0,36	0,36	46,1	0,64	0,64	0,567	0,01	0,567	0,675	0,675	86,4	R1
P2	0,331	0,691	42,3	0,0538	0,694	0,0473	0,01	0,614	0,0628	0,738	8,03	R1
P3	0,126	0,817	16,1	0,0186	0,713	0,042	0,01	0,656	0,0229	0,761	2,93	R1
P4	0,0838	0,9	10,7	0,00104	0,714	0,0129	0,01	0,669	0,004	0,765	0,512	R1
Orthogonal in X(OPLS)		0,061			0,000608							
O1	0,061	0,061	7,81	0,000608	0,000608							R1
Orthogonal in X(PCA)		0,0349			0							
O2	0,0136	0,0136	1,74	0	0							R1
O3	0,00913	0,0228	1,17	0	0							R1
O4	0,00372	0,0265	0,477	0	0							R1
O5	0,00313	0,0296	0,401	0	0							R1
O6	0,00196	0,0316	0,252	0	0							R1
O7	0,00111	0,0327	0,142	0	0							R1
O8	0,000813	0,0335	0,104	0	0							R1
O9	0,00074	0,0342	0,0947	0	0							R1
O10	0,000647	0,0349	0,0829	0	0							R1
Orthogonal in Y(OPLS)										0,228		
O1									0,14	0,14	17,9	R1
O2									0,0879	0,228	11,3	R1
Orthogonal in Y(PCA)										0,00631		
O3									0,00239	0,00239	0,306	R1
O4									0,0015	0,00389	0,192	R1
O5									0,00116	0,00505	0,148	R1
O6									0,000592	0,00564	0,0758	R1
O7									0,000465	0,00611	0,0595	R1
O8									0,000209	0,00631	0,0268	R1

Figure 6.14: Model window of the O2PLS model for multiple spectral datasets.

6.5.1 Analyzing the joint variation in the spectral data

The O2PLS model shows that 90% of the variation in the NIR/IR block correlates with 76% of the variation in the Raman block, which is a substantial overlap. By plotting the score vectors of the largest components, the structure of this joint information can be evaluated (no plots shown). The conclusion is that spectra from the three analytical platforms have many features in common and provide important information about the investigated samples.

6.5.2 Analyzing the unique variation in the spectral data

In addition to estimating the joint variation, the O2PLS model indicates that 9.6% of the variability is unique to the combined NIR/IR block and that 23.4% of the variability is unique to the Raman block. We will use the drill-down capability of the SIMCA software to analyze the first part more closely. Figure 6.15 shows how the unique NIR/IR variability evolves with time across the 128 samples. The total time span is five days.

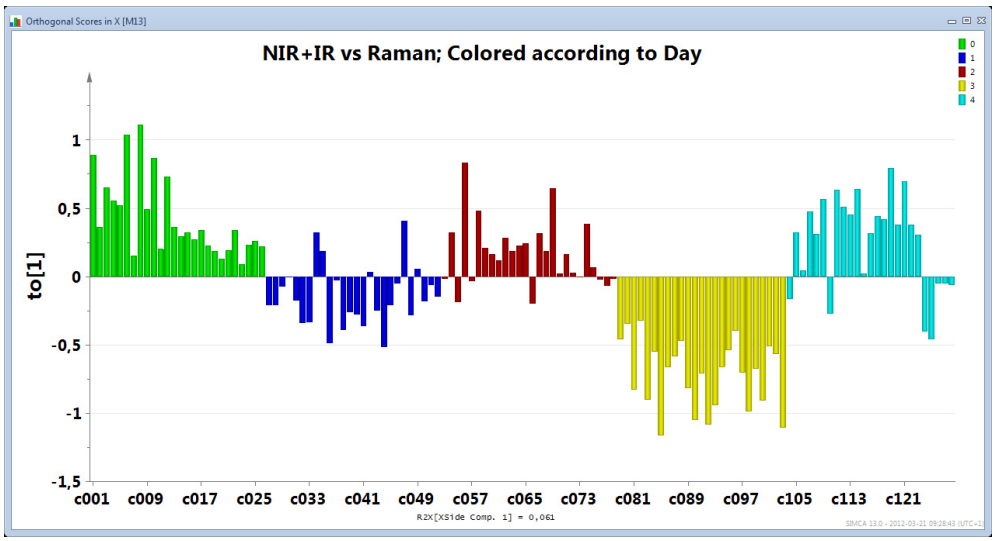


Figure 6.15: Column plot showing the distribution of the 128 samples in the NIR/IR unique feature space. The time span is five days and there is a systematic deviation for the samples measured during day four. This deviation is not found in the Raman spectra. The plot is colored according to measurement day.

There is a weak trend during the first two days of the measurement campaign, resulting in slowly decreasing score values along the component. Subsequently, when moving from day three to day four, there is a sharp jump, and then a similar, but opposing, jump when going from day four to day five. A detailed analysis of the unique variation (Figure 6.16) suggests this day-to-day shift is linked to varying moisture content in the samples. NIR and IR measurements are more affected by such variations than Raman spectroscopy.

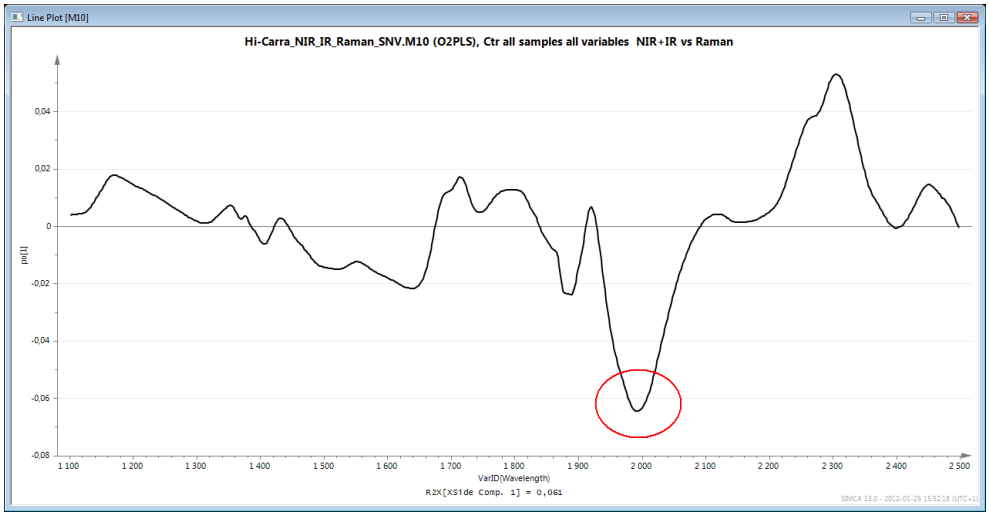


Figure 6.16: NIR-loading spectrum corresponding to Figure 6.15. This plot demonstrates that the region around 1970 nm strongly contributes to the unique variation. This suggests that the systematic shift in sample properties during day four is connected to a moisture problem. NIR and IR measurements are more affected by such variations than Raman spectroscopy.

6.5.3 Summary of the multiple spectral datasets example

In this work, O2PLS was used to detect subtle changes in sample properties over time. The changes in the sample properties were determined to arise from a moisture problem. NIR and IR measurements are more affected by such variations than Raman spectroscopy. This day-to-day change in sample properties was not noticed when using PLS as the modeling basis. Hence, O2PLS helped with unraveling and interpreting previously unknown sample differences. Knowing, understanding and removing such properties will aid in the long-term

fine-tuning of the measurement platforms. A more detailed analysis of this example is given in Chapter 19, which deals with hierarchical modeling.

6.6 Questions for Chapter 6

1. What is the objective of O2PLS?
2. What is meant by joint variability?
3. What is meant by unique variability?
4. Which two methods are used to estimate the components expressing unique variability?

6.7 Summary and discussion

There is an increasing variety of analytical and imaging methods that are used for characterizing complex samples and manufacturing processes. Such analytical and process data provide a very satisfactory basis for data mining and data integration because, although the data are numerous, they are often of high precision and the rewards are substantial. The O2PLS analysis of analytical and process data provides models that are interpretable and transparent, since it is possible to separate joint and unique variabilities from each other.

It should be emphasized that O2PLS has an operational focus different from PLS and OPLS. The objective of PLS and OPLS is regression between two blocks of data. The objective of O2PLS is integration of two or more blocks data. To emphasize this difference in modeling objectives, we use different terminology. Whereas with OPLS it is practical to distinguish between *predictive* and *orthogonal* sources of variability, when using O2PLS it becomes natural to consider *joint* and *unique* sources of variability. The joint variation represents the information overlap between X and Y, i.e., the systematic structure that is present in both X and Y. Any unique variation corresponds to information beyond the overlap and is exclusively found in either the X- or the Y-block.

By means of the two examples presented in Sections 6.4 and 6.5, we have shown the analytical capability of O2PLS to overview complicated real-world multi-block problems in terms of transparent models where every component has a clear meaning simplifying its interpretation. There is no doubt that through O2PLS it was possible to uncover more useful information than would have been possible through the corresponding PCA and PLS models.

Recently, the O2PLS method was extended to OnPLS [Löfstedt and Trygg, 2011] to cover any number of datablocks. The proposed method is equivalent to O2PLS in cases involving two matrices, but generalizes to cases involving many matrices without giving preferences to any particular matrix. Being fully symmetric, OnPLS extracts a minimal number of globally predictive components that exhibit maximal covariance and correlation. Furthermore, the method allows to study unique variation, i.e. local phenomena captured in the data that are specific to individual combinations of matrices or to individual matrices.

It is anticipated that in the near future OnPLS will find rapid applicability and widespread use in process industry to tackle problems involving integrating data arising from multiple measurement platforms or process data that are stored in multiple databases. One such principal problem involves following data in a multi-stage manufacturing process and finding out how the information structure is preserved, or, alternatively, altered, when processing the product-to-be across the various production stages. Such information structure changes will indicate if the entire production chain is robust or partly unstable. Another conceivable main type of problem involves scale-up, from lab, to small pilot, to large pilot, to production, and finding out which production features are in common and which differ across scales.

Other extensions of the core O2PLS methodology involve OnPLS algorithms addressing the bi-model three-way matrix structures often seen in the analysis of food sensory and consumer data [Löfstedt, et al., 2012a], and problems consisting of chained matrices, so called path models [Löfstedt, et al., 2012b].

7 Multivariate characterization

7.1 Objective

The objective of Chapter 7 is to discuss multivariate characterization, which is an important application area of PCA. Multivariate characterization consists of three steps:

(1) measurement of relevant multivariate data on a set of observations,
(2) creation of a low-dimensional map from these data using PCA, and
(3) use of the resulting map to select a smaller set of observations that is *representative* of the whole set.

Multivariate characterization is illustrated by two examples, the first dealing with the selection of suitable solvents for organic synthesis, and the second focusing on the selection of representative non-ionic surfactants for structure-activity studies.

7.2 Introduction

Representative experiments are crucial for achieving reliable results in experimental work, regardless of whether the experiments take place in the laboratory, the pilot plant, or full scale production. Such representative experiments are easily obtained by DOE when factors are continuous and controllable, like pH, temperature, or pressure. With quantitative factors, experimental designs might be laid out according to factorial or fractional factorial designs [Box, 1978].

However, research, development, and production, rarely involve the study of quantitative factors only. On the contrary, there are often other types of factors which are relevant, but which cannot be manipulated directly according to DOE. Consider, for example, the variability in a manufacturing process induced by different batches of raw material, or of a catalyst. Because of the discrete variation from one batch to another, it becomes difficult to ascribe a single numerical value to each batch, indicating its quality, composition, color, viscosity, etc., in relation to the other batches. Consequently, it will be difficult to introduce changes in such discrete factors, as would have been proposed by a statistical experimental design.

It is in this perspective that multivariate characterization is useful. Multivariate characterization is a method of *quantifying discrete changes*, changes which might occur between different batches of raw material, catalysts, solvents, compounds in drug design, and other "cases". This quantification is based on approximating those latent factors which represent the important properties of the studied "cases", by means of principal component analysis. An interesting and convincing example of how multivariate characterization can be used to map properties of chemically modified celluloses, when batch-wise variation is present, was presented by Svensson and coworkers [Svensson, *et al.*, 1997]. In other applications, Börås *et al.* and Elg-Kristoffersson *et al.* applied multivariate characterization to describe properties of pulp fibers in paper industry [Börås, *et al.*, 1997; Elg-Kristoffersson, *et al.*, 2002]. Other interesting studies involve application of multivariate characterization to beer brewing [Duarte, *et al.*, 2002], classification of activated carbon adsorption filters [Wu, *et al.*, 2003] and proteomics data analysis [Long, *et al.*, 2004].

Multivariate characterization is of growing importance in food, dairy and other industrial sectors [Güler, 2007; Piazza, *et al.*, 2007; Rudnitskaya, *et al.*, 2007; Dumarey, *et al.*, 2011; Chorell, *et al.*, 2012; Mogalaka-Matlala, *et al.*, 2013]. Section 18.4.4 provides an example

of raw material characterization in conjunction with a process analytical technology (PAT) application.

In the next few sections we will outline the main steps of multivariate characterization, and use two datasets as illustrations of these stages. The discussion will be focused toward applying multivariate characterization when the observations are compounds, or blends of related compounds. The application to batch-wise variation among raw materials is obvious.

7.3 Rationale of multivariate characterization

Whenever we wish to quantify changes in a discrete multi-level factor, multivariate characterization, based on PCA is the method of choice. Multivariate characterization has been used successfully in several areas for selecting sets of representative compounds and substituents, e.g., in synthetic organic chemistry [Nordahl and Carlson, 1993; Carlson & Carlson, 2005], combinatorial chemistry [Lundstedt, *et al.*, 1997], medicinal chemistry [Winiwarter, *et al.*, 1998; Giraud, *et al.*, 2000; Eriksson, *et al.*, 2004], environmental chemistry [Tysklind, *et al.*, 1995; Ramos, *et al.*, 1997], QSAR [Eriksson and Hermens, 1995a; Hellberg, 1986; Sandberg, *et al.*, 1998], and microbiology [Marvanova, *et al*., 2001]. Similarly, multivariate characterization has been used to select stationary phases in chromatographic applications (TLC, HPLC, GC), and for design of experiments (DOE) in cases where variability is caused by changes between batches of raw material [Börås, *et al.*, 1997; Svensson, *et al.*, 1997].

Variation in the levels of a discrete factor often involves changes of properties on a molecular scale. The problem is, however, that a discrete change from, say, solvent A to solvent B, will change a number of molecular properties, not just one isolated property. A change of a substituent in one part of a molecule or a polymer will alter several properties, e.g., electron distribution, lipophilicity, chemical reactivity, steric configuration, and hydrogen bond donor and acceptor abilities. It is these intrinsic, latent molecular properties which govern the chemical behavior of the molecule or the polymer.

However, changes in intrinsic molecular properties cannot easily be measured directly. What can be measured are macroscopic properties, which are manifestations of the intrinsic properties. Macroscopic information can be acquired through the use of relevant chemical, physical, and perhaps even biological model systems. Fortunately, the instrumental revolution in analytical and bio-analytical chemistry has equipped us with a wide variety of measurement techniques. For instance, spectroscopic (IR, NIR, NMR,…) and chromatographic (TLC, HPLC, GC, …) methods have found widespread applicability in multivariate characterization, thereby supplementing the more conventional molecular descriptors related to physical (boiling point, melting point, refractive index, …), chemical (ionization potential, lipophilicity, reactivity, …), and structural (bond lengths, bond angles, van der Waals radii, ...) properties of molecules.

For a given problem, it is important to identify model systems (variables) which are relevant for the application. These model systems are utilized to derive a battery of descriptor variables, which map the intrinsic properties of a class of chemically and functionally related substances. Within a class of related substances, the compounds (observations) are in some respect similar to each other, and it is therefore reasonable to assume that a gradual variation of measured chemical properties corresponds to an analogously gradual variation of the intrinsic molecular properties. Furthermore, it is reasonable to expect that properties which are measurable, and which depend on the same underlying intrinsic property, will be more or less correlated to each other. Hence, many correlated variables in a multitude of chemical and structural descriptors are not uncommon.

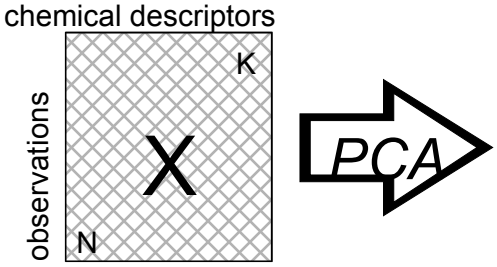
We know from the earlier chapters that correlated variables may beneficially be analyzed by multivariate projection techniques. The application of PCA to this kind of data will produce new variables, the PCA scores, which adequately summarize the intrinsic molecular properties. Hence, these PCA scores are often termed *principal properties*, or simply PPs [Hellberg 1986]. Such PPs are invaluable for selecting a representative and diverse sub-set from a series of chemical compounds, or for selecting representative starting materials when batch-wise variation is present.

7.4 Main steps of multivariate characterization

There are three main steps in multivariate characterization, namely (i) compilation of a battery of chemical descriptors, (ii) derivation of principal properties, and (iii) selection of a representative and diverse set of observations. We will now briefly review the essential aspects of these three steps. The illustration that we have selected is called SOLVENT and deals with how to choose representative solvents for synthetic organic chemistry. Professor Rolf Carlson at the University of Tromsö, Norway, is the originator of this application.

7.4.1 Compilation of chemical descriptors (Example: SOLVENT)

In order to adequately portray the intrinsic properties of a series of compounds, one must first gather chemical descriptors. These descriptors may arise from experimental model systems, mostly chemical and physical in nature, but they may also be derived from theoretical considerations, for example, by using quantum-chemical molecular orbital calculations. It is, of course, difficult to know in advance exactly which chemical descriptors will be useful, but experience from the fields of QSAR and QSPR shows that descriptors related to chemical reactivity, lipophilicity, polarizability, polarity, electron distribution properties, etc., provide a good starting basis. The collected descriptors form the X-matrix (Figure 7.1).



	ObsNum	ObsName	t[1]	t[2]
7	1	W01_WATER	-4.8815	-1.1908
8	2	G02_Formamide	-4.9407	1.2276
9	3	A03_1.2-Ethanediol__	-3.5664	0.45437
10	4	A04_Methanol	-2.7109	-2.7815
11	5	G05_N-Methyl_formami	-5.1221	1.3178
12	6	A06_diethylenglycol	-2.6615	1.3319
13	7	A07_triethylenglycol	-3.4296	1.4872
14	8	A08_2-Methoxyethanol	-2.2226	-1.3235
15	9	G10_N-Methylacetamid	-5.2038	1.8804
16	10	A11_Ethanol	-1.949	-2.5121
17	11	A12_2-AminoEthanol	-3.0024	0.59628
18	12	F13_Acetic_acid	-2.0907	-0.46001
19	13	A15 Benzylalcohol	-0.99517	1.8138

Figure 7.1: The first step of multivariate characterization involves a multitude of chemical descriptors to encode the intrinsic molecular properties within a set of chemical compounds. These intrinsic properties are then estimated mathematically by the computation of the so called principal properties, PPs, of the compounds. The PPs correspond to a set of scores derived by PCA.

In the SOLVENT application, the question was one of identifying a set of representative solvents for synthetic organic chemistry. The assumption was that intrinsic properties of the studied solvents, could, at least partly, be seen in well chosen chemical model systems. Rolf Carlson and coworkers brought together a series of 103 solvents, by browsing through different inventories of chemical substances, such as, the CRC Handbook of Chemistry and Physics. In order to account for the intrinsic solvent properties, a set of nine chemical property values were collected. These were: (1) melting point; (2) boiling point; (3) dielectric constant; (4) dipole moment; (5) refractive index; (6) ET30, a transition energy from UV-spectroscopy of an organic dye; (7) density; (8) log P, octanol-water partition coefficient; (9) water solubility.

7.4.2 Derivation of principal properties (Example: SOLVENT)

The second main step in multivariate characterization involves a PC analysis of the compiled descriptor data (Figure 7.1). This results in PCA scores (= principal properties) of a class of compounds. It should be noted that many of the chemical descriptors might be correlated. This redundancy, in the sense that many measured variables reflect a certain intrinsic molecular feature, will tend to stabilize the estimates of the PPs in much the same way as an ordinary arithmetic average is improved by repeated measurements.

When PCA was applied to the 103 (solvents) by 9 (chemical descriptors) SOLVENT dataset, a two-component model resulted. This model had an R^2X of 0.68 and a Q^2X of 0.46 (Figure 7.2). The first component explains 40% of the variation and the second 28%. This suggests that the nine chemical descriptors encode two intrinsic molecular properties. In addition to the overall R^2X and Q^2X-values, these can be computed also for the individual chemical descriptors (see plot in Figure 7.3).

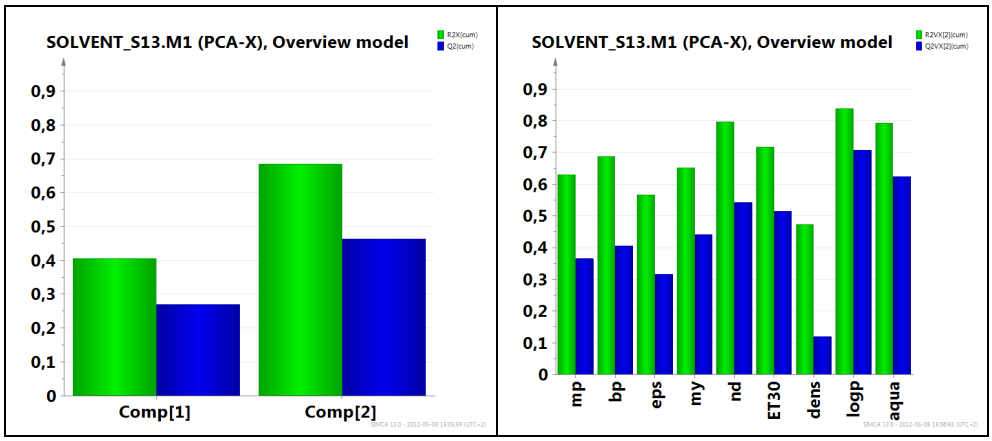


Figure 7.2: (left) Model overview of the PCA model of the SOLVENT dataset. After two components R^2X amounts to 0.68 and Q^2X to 0.46. Figure 7.3: (right) Individual R^2X and Q^2X-values of the nine chemical descriptors. The descriptors are: (1) melting point; (2) boiling point; (3) dielectric constant; (4) dipole moment; (5) refractive index; (6) ET30, a transition energy from UV-spectroscopy of an organic dye; (7) density; (8) log P, octanol-water partition coefficient; (9) water solubility.

Figure 7.3 demonstrates that two chemical descriptors, log P and water solubility (aqua), are well described and predicted by the PCA model. Q^2X for these variables exceeds 0.6. In a second group of variables, we find the six descriptors melting point, boiling point, dielectric constant, dipole moment, refractive index, and ET30, for which Q^2X ranges between 0.30 and 0.55. Hence, these are rather well explained and predicted by the PCA model. Finally, we have density, which has a near-zero Q^2X (< 0.15) and rather low R^2X (≈ 0.45). This variable is not well explained or predicted.

The two PCA scores are called the PPs of the 103 solvents. Figure 7.4 shows a scatter plot of PP2 versus PP1 (t_2 vs. t_1), in which each point represents one solvent. The distribution of solvents is rather uniform, indicating a homogenous class of compounds. There is no need to discard any strongly outlying solvent. There are four observations positioned outside the elliptical tolerance area, but with 103 observations, around five are expected to show up outside this limit at the 5% level.

In order to ascribe a physico-chemical meaning to the two PPs, we look at the p_1/p_2 loading plot displayed in Figure 7.5. The first component is dominated by lipophilicity (log P), water solubility, and the UV-related variable ET30. The investigators interpreted the first PP as mirroring mainly water solubility. The second component is influenced by boiling point, refractive index, and melting point, and was hence interpreted as reflecting polarizability.

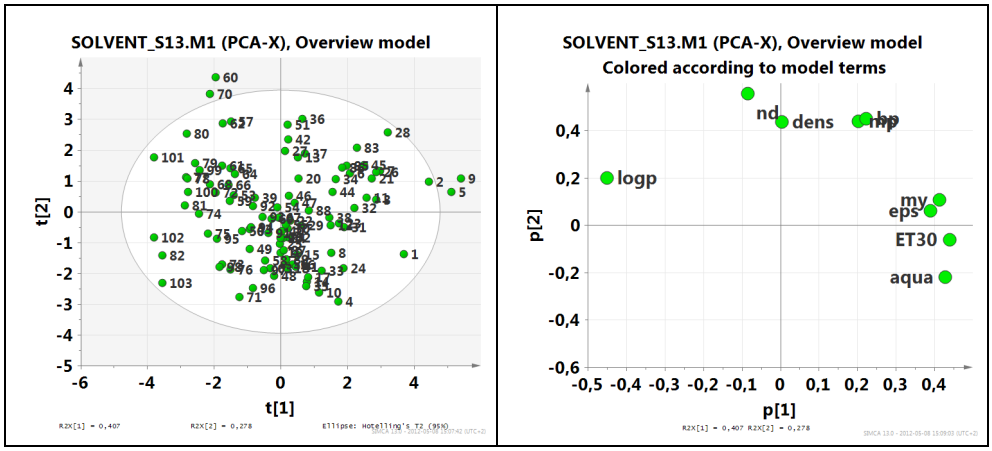


Figure 7.4: (left) Scatter plot of solvent PP2 against PP1. Each plot mark denotes one solvent. Figure 7.5: (right) PCA loading plot with the second loading vector p_2 plotted against the first loading vector p_1. Based on the distribution of the nine chemical descriptors, the PPs were tentatively interpreted as reflecting water solubility (PP1) and polarizability (PP2).

7.4.3 Selection of a representative set of observations (Example: SOLVENT)

The interesting feature of the PPs is that they are mathematically independent of each other (orthogonal) and limited in number (often between two and four). This implies that they may well be used in statistical experimental design schemes, such as fractional factorial designs [Box, *et al.*, 1978] or D-optimal designs [De Aguiar, *et al.*, 1995], providing a representative and diverse set of compounds. This process is generally called *multivariate design* or *statistical molecular design* in the pharmaceutical industry [Linusson, *et al.*, 2001; Gabrielsson, *et al.*, 2002; Wold, *et al.*, 2004; Chorell, *et al.*, 2012]. Any set of compounds selected in accordance with a multivariate design will automatically have a good chemical diversity and spread in the latent variables (PPs). This, in turn, implies a lesser degree of multicollinearity in the data, which is beneficial for the subsequent data analysis.

In a fractional factorial design, all PPs are usually explored at two levels, one low (called the minus level) and one high (termed the plus level), but there may be as many as five levels [Eriksson and Johansson, 1996]. Additionally, one or several "center-points" are usually included in the designs. The PPs are varied simultaneously over the entire training set, implying that each individual PP will be explored at all levels of the others (Figure 7.6). This procedure will yield a table of plus and minus signs specifying the desirable PP-combinations, i.e., the optimal combinations of chemical properties required for the set of representative and diverse substances.

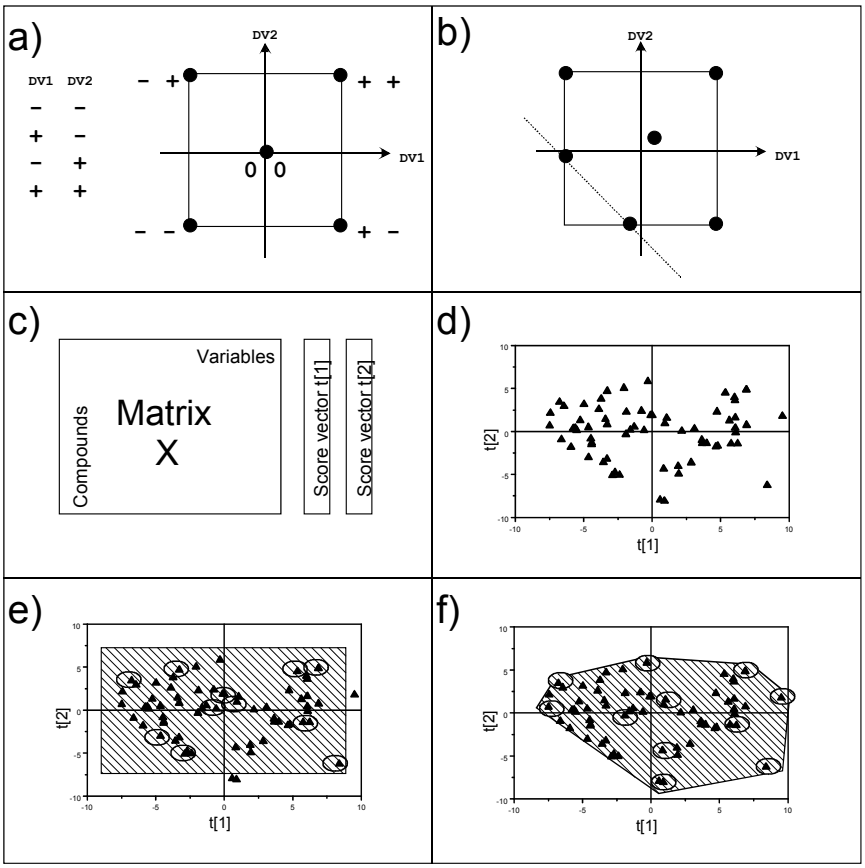


Figure 7.6: An illustration of multivariate design. (a) A regular factorial design in two design variables (PPs), encoding the selection of four corner experiments and one center experiment (0 0). (b) A D-optimal design laid out in the two design variables, which is better adapted for covering an irregular experimental domain. The irregularity is created by the constraint depicted by the dotted line. (c) PCA of chemical data giving two PPs (here t_1 and t_2). (d) The PPs are orthogonal and may be used as design variables. The projected compounds define an irregular domain ("experimental region"). (e) Multivariate design based on a factorial design, replicated once. Two compounds are selected from each quadrant, and two "center"-compounds from the interior part of the score space. (f) Multivariate design based on a D-optimal design. With the D-optimal design compounds are selected which are as close to the extreme vertices as possible. With the factorial design compounds are sought as close to the quadrant corners as possible. Both design types are favorably augmented with some interior points.

A design laid out according to a fractional factorial design (FFD) exhibits a regular geometry (Figure 7.6). As a consequence, an FFD may occasionally dictate a PP-combination that does not exist in nature, or is not attainable synthetically. Thus, there may exist constraints in the space defined by the PPs, rendering the regular geometry of an FFD unattainable. Such constrained situations are better tackled using D-optimal designs, which are more general and flexible by virtue of their working directly with the numerical values of the PPs [Baroni, *et al.*, 1993a]. Hence, in the context of systematically selecting representative compounds, a D-optimal design is often more practical than a fractional factorial design, because it is possible to reduce the number of compounds, to exclude chemical structures too difficult to test, and to include certain "mandatory" compounds tested for reference purposes.

The D-optimality approach usually needs a computer to generate the design, in contrast to a fractional factorial design which may be created by hand. The D-optimal design is a selection of experiments (in multivariate design often compounds or substituents) that maximizes the volume spanned in the PP-space. This corresponds to maximizing the information content of the chosen compounds.

We now consider the SOLVENT example. Based on the derived PPs, different selection strategies may be attempted. One approach is based on a selection of solvents that gives a rather uniform coverage of the PP-space (Figure 7.7). Such a selection can be obtained by identifying solvents that form a nearly regular lattice according to a factorial design.

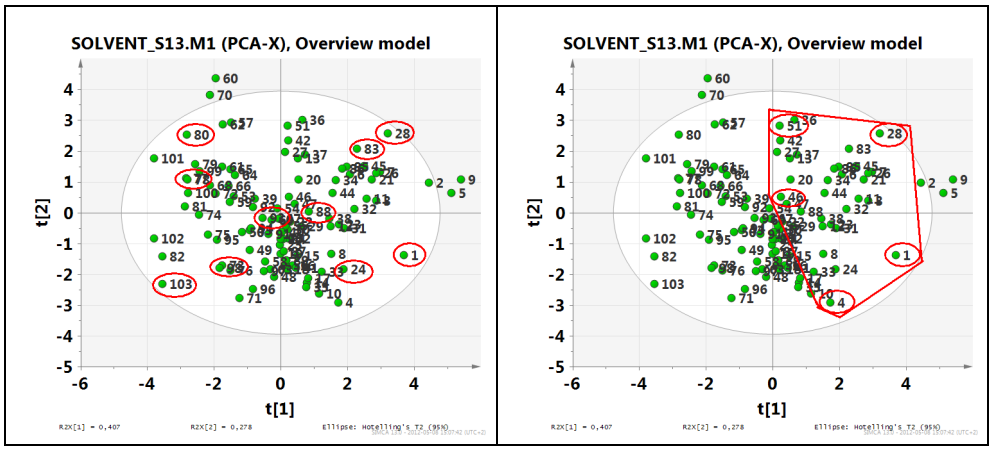


Figure 7.7: (left) One solvent selection strategy is to select solvents that spread evenly in the PP-space. This can be accomplished by a fractional factorial design. Figure 7.8: (right) Another selection approach is to restrict the choice of solvents to a certain type of chemical functionality or structure. This is normally accomplished by a local multivariate design tailored to map a constrained region of the PP-space.

The first selection principle (Figure 7.7) provides a series of solvents exhibiting a large degree of chemical diversity. This makes it possible to conduct an initial screening of the impact of solvent on the investigated system (e.g., a chemical synthesis).

However, in many cases drastic changes of solvent properties may be undesirable, and the situation sketched in Figure 7.7 will not apply. Perhaps only a limited PP-area is relevant. The latter case is illustrated in Figure 7.8, where a local multivariate design has been constrained to a small region of the PP-space. How to handle constrained principal property spaces in drug design is exemplified in the chapter on lead finding and lead optimization (see supplementary book material available on www.umetrics.com).

A restricted PP-region corresponds to considering solvents with limited variation in intrinsic properties. In Figure 7.8, the solid lines indicate the area where alcohols (ethanol, 1-propanol, 2-butanol, etc.) are encountered. A more thorough discussion concerning the application of multivariate design when data are clustered in PP-space is given in the chapter on QSAR (see supplementary book material available on www.umetrics.com), where we describe an approach to cluster-based design in environmental QSAR.

Obviously, the choice of observation selection strategy will vary from application to application. The strategies suggested above will enable a systematic exploration of the entire PP-space or of part of it, by selecting representative "cases", which ensure that the intrinsic properties of a set of observations are spanned appropriately.

7.4.4 A short discussion

The basic idea of multivariate characterization is thus to formulate a quantitative numerical representation of a qualitative, discrete change. This is accomplished by principal component analysis of the collected data. The benefit of the PPs is that they comprise a good basis for the selection of representative "cases".

We have earlier stressed the need to use several descriptors in multivariate characterization in order to adequately describe the available compounds. By "several" we envision at least 10-20 descriptors. These variables may be measurable in chemical model systems, calculable using quantum-chemical orbital theory, or simply based on atom and/or fragment counts. It is often difficult to know beforehand which chemical descriptors will be found useful however, and an adequate answer will usually not be obtained until verifying experiments (predictions for new observations) are carried out.

Finally, we recognize that the number of descriptors reflecting certain intrinsic molecular properties determines the order of emergence of the PPs since the first PC explains the largest part of the variation in the data, the second PC the second largest source of variation, and so on. Where the optimal "weights" used by nature are unknown, it may be practical to scale the input descriptors to unit variance prior to computing the principal properties. Alternative approaches might involve more elaborate scalings to up-weight or down-weight certain chemical descriptors, thereby changing the order and significance of the PPs.

7.5 Main example: SURFACTANTS

In this section, we shall consider a study where multivariate characterization, design, and modeling were applied to a class of non-ionic surfactants. Non-ionic surfactants, here, ethoxylated fatty alcohols, are important constituents of detergent blends. Because the manufacturing process usually involves a catalytic polymerization of ethylene oxide (EO) units on a fatty alcohol, a distribution of EO chain lengths is obtained. This means that in QSAR or QSPR one must characterize the technical blend, and not a single compound. This characterization of technical blends makes this application unique.

Lindgren and coworkers [Lindgren, *et al.* 1996; Sjöström, *et al.*, 1997] investigated a series of 38 non-ionic EO-based surfactants. The objective was to achieve a good balance between washing performance and environmental performance. The general structure of these surfactants was $(EO)_n - R$, where R is the hydrophobic part. R may contain straight carbon chains, branched chains, aromatic moieties, unsaturation, etc. For such surfactants there has always been a trade-off between washing performance and biodegradation. Washing performance means the ability to remove dirt from soiled material. Many of the intrinsic surfactant properties are affected by the properties of R. For instance, branching gives good washing performance but poor biodegradability, whereas straight chains have the reversed pattern. Thus, it appears that a useful compromise might be a straight chain with some degree of branching. This was the working hypothesis of this study.

The questions asked by the researchers were: (1) Is it possible to quantitatively model the performance of technical blends as a function of chemical properties? (2) Which representative surfactants should be studied?

The experimental work and multivariate data analysis were carried out in six steps:

Step 1: Multivariate characterization of technical surfactant blends
Step 2: Training set selection
Step 3: Detergency measurements
Step 4: Toxicity measurements
Step 5: PLS regression between chemical descriptors (X) and detergency/toxicity (Y)
Step 6: Predictions of response profiles (Y) for non-tested surfactants

7.5.1 Multivariate characterization of technical surfactant blends

To account for the chemical properties of the non-ionic surfactant blends, 19 chemical descriptors were used. These descriptors reflect lipophilicity (log P), molecular weight, branching of R, technical blend properties, critical micellar concentration, and some indices describing the hydrophobic/hydrophilic balance in a surfactant molecule. Perhaps the most innovative descriptors were established by using chromatography (HPLC) to describe the technical blends. From the resulting chromatogram, the maximum peak (maxEO) related to moles EO per mole fatty alcohol was determined. To describe the shape of this EO-distribution peak, the peak widths at 33% peak height (w33EO) and 66% peak height (w66EO) were registered. In total, N = 38 surfactant blends were investigated.

A PCA of the 38 x 19 data matrix gave a four-component model with $R^2X = 0.86$ and $Q^2X = 0.54$. The PCA scores of the two first components are plotted in Figure 7.9, and the corresponding loadings in Figure 7.10. The last two PPs only account for 24% of the variation, and are therefore not considered in the current interpretation of the model.

Based on an inspection of the loadings (Figure 7.10) the first PP was related to surfactant lipophilicity. This is because high values for PP1 (high lipophilicity) correspond to high values for log P, ethylene oxide (EO) content, molecular weight (Mw), and cloud point (CP), and low values for the amount of non-ethoxylated fatty alcohol (F-alcohol), the critical packing parameter (CPP) and the reduced packing parameter (redCPP). The second PP was understood as reflecting the hydrophobic/hydrophilic balance of the surfactants. It was discovered that high values of PP2 (low hydrophobic/hydrophilic balance) correspond to low values of the two indices reflecting hydrophobic/hydrophilic balance (Davis and Griffin) and high value of redC/C, i.e., a branched hydrophobic part.

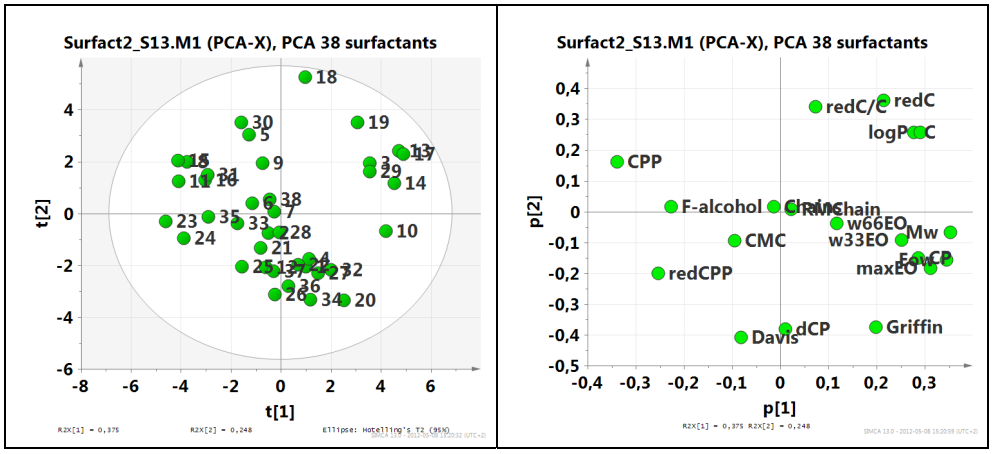


Figure 7.9: (left) PCA t_1/t_2 score plot of the model for the SURFACTANT dataset. This score plot accounts for 62% of the variation in the 19 chemical descriptors. Surfactants found in the upper right-hand corner are too lipophilic to be of interest for further studies. Figure 7.10: (right) PCA p_1/p_2 loading plot corresponding to the scores shown in Figure 7.9. The importance of log P and log P-related descriptors is obvious.

The distribution of surfactants in Figure 7.9 shows two strong clusters. Since the smaller cluster, located in the upper right-hand part, comprises surfactants that are too lipophilic for the experimental objective, the eight surfactants belonging to this cluster were excluded. When a surfactant is very lipophilic it might be difficult to maintain the desired balance between good washability and biodegradation within a reasonable time.

Thus, a new PCA was computed on the remaining 30 detergents. This model displayed the performance statistics $R^2X = 0.76$ and $Q^2X = 0.52$, after three components. The first PP explains 38%, the second 22%, and the third 16% of the total variation. The updated PPs are plotted in Figures 7.11 (t_1/t_2) and 7.12 (t_1/t_3).

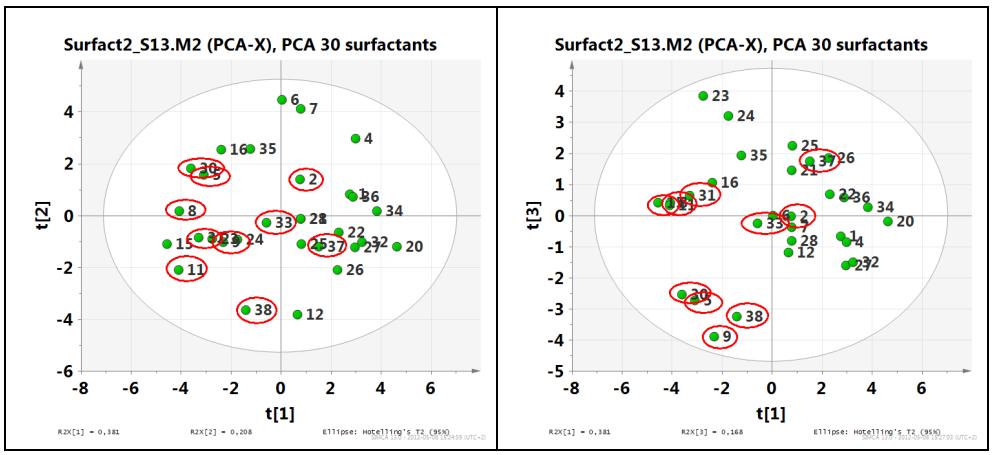


Figure 7.11: (left) Scatter plot the updated PP2 versus PP1 for the reduced set of surfactants. Each number corresponds to one surfactant. The location of the ten surfactants chosen is indicated by ellipses. These surfactants form the so called training set, that is, the set of surfactants on which to train (calibrate) the final PLS model. Figure 7.12: (right) Scatter plot of the updated PP3 versus PP1.

7.5.2 Training set selection

Subsequently, the three updated PPs were used for multivariate design. The authors did not explicitly use any conventional statistical experimental design (factorial, fractional factorial, D-optimal,...), but identified ten representative non-ionic surfactants providing a reasonable coverage of a restricted PP-area. The location of the ten surfactants is indicated in Figures 7.11 and 7.12 by ellipses. These surfactants form the so called *training set*, that is, the set of surfactants on which to *train* (calibrate) the PLS model. This PLS model will connect information in the chemical descriptors (X) to information concerning washing performance and acute toxicity (Y).

7.5.3 Detergency measurements

In the third step, the training set of surfactants was subjected to washing performance testing. The authors used response surface modeling [Box, *et al.*, 1978] to help describe the washing performance. The washing efficiency was summarized as the performance after 15 minutes washing. Three response variables were measured, (1) % removed soil (^{13}C labeled triolein, YDet), (2) optimal concentration of surfactant (YConc), and (3) optimal washing temperature (YTemp). These responses are summarized in Table 7.1. More experimental details are found in the original articles [Lindgren *et al.*, 1996; Sjöström, *et al.*, 1997].

7.5.4 Acute toxicity measurements

In the fourth step, the training set was tested for acute toxicity toward a fairy shrimp, so called Toxkit measurements. The experimental results were summarized as an LC_{50}-value (concentration causing 50% lethality among the test organisms). Table 7.1 lists the log LC_{50}-values (YTox).

Table 7.1: Summary of detergency performance and acute toxicity.

Surfact.	YDet	YConc	YTemp	YTox
2	86.3	1.4	60	1.29
5	81.8	1.6	48	0.52
8	82.4	1.9	44	0.96
9	81.8	1.5	82	0.55
11	85.4	2.25	54.5	1.25
30	78.8	1.7	48	0.59
31	84.6	2.8	50	1.77
33	88.9	1	74	1.11
37	85.9	2	67	1.52
38	86.5	1.1	61	

At this stage, we make a few remarks concerning the data in Table 7.1. The surfactant numbering accords with the original publication. YDet represents % removed soil from soiled cloth. This is an important response, which should be maximized. YConc corresponds to the concentration (mM) of surfactant required for good washing performance. This response is regarded as of minor importance, but it should, if possible, be kept low. YTemp is the washing temperature (°C) required for good washing performance. This is an important response. It should be kept low, preferably in the range 40-60 °C. Finally, YTox depicts the acute toxicity, that is, log LC_{50} (μM). This is also an important response, which should be maximized (high numerical values correspond to low toxicity). Also observe that there is one missing data value in this last response. In summary, it appears that surfactants no. 2 or 37 might be promising candidates for fulfilling the experimental objectives.

7.5.5 PLS regression between chemical descriptors (X) and detergency/toxicity (Y)

The relationship between the four responses (YDet, YConc, YTemp & YTox) and the 19 chemical descriptors was modeled by PLS. Three PLS components were obtained, and the cumulative R^2Y- and Q^2Y-values are plotted in Figure 7.13. After three components, R^2Y amounts to 0.78 and Q^2Y to 0.46, which are excellent values considering that technical surfactant blends are investigated, and not pure compounds. Figure 7.14 shows R^2Y and Q^2Y for each response. Evidently, all responses are modeled and predicted well, with Q^2Y´s ranging between 0.40 and 0.55.

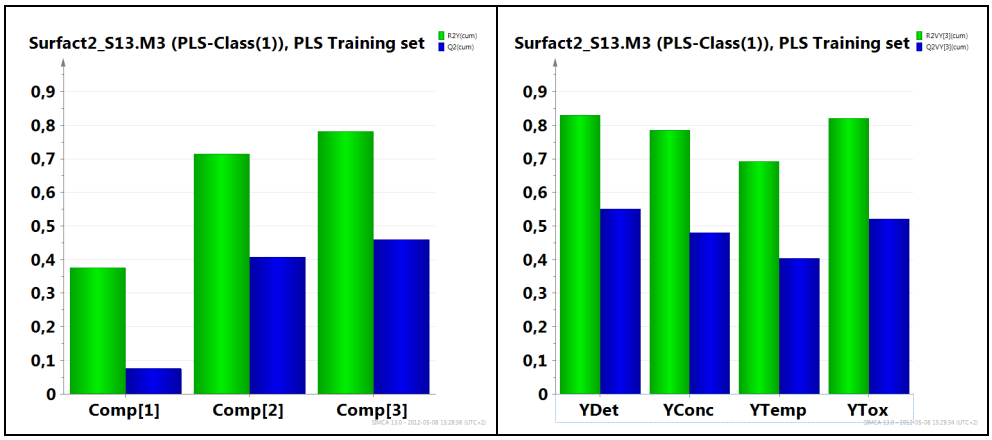


Figure 7.13: (left) Model overview of the PLS model based on the selected set of 10 surfactants. The three-component model explains 78% (R^2Y = 0.78) and predicts 46% (Q^2Y = 0.46) of the variation in the four response variables. Figure 7.14: (right) Plot of cumulative R^2Y and Q^2Y of each response. The individual Q^2Y-values are 0.55 (YDet), 0.48 (YConc), 0.40 (YTemp), and 0.52 (YTox).

In order to interpret the PLS model we shall consider its scores and weights. With three PLS components, there are three PLS score plots of the t/u-type to explore. Figures 7.15 to 7.17 show these score plots. A close inspection of the t_1/u_1 score plot (Figure 7.15) shows a mild curvature in the relation between X and Y. Hence, the second model component (t_2/u_2; Figure 7.16) represents a compensation for non-linearity. In the last component (t_3/u_3; Figure 7.17) the inner relation is linear. The conclusion is that there exists a strong correlation structure between the 19 chemical descriptors and the four response variables.

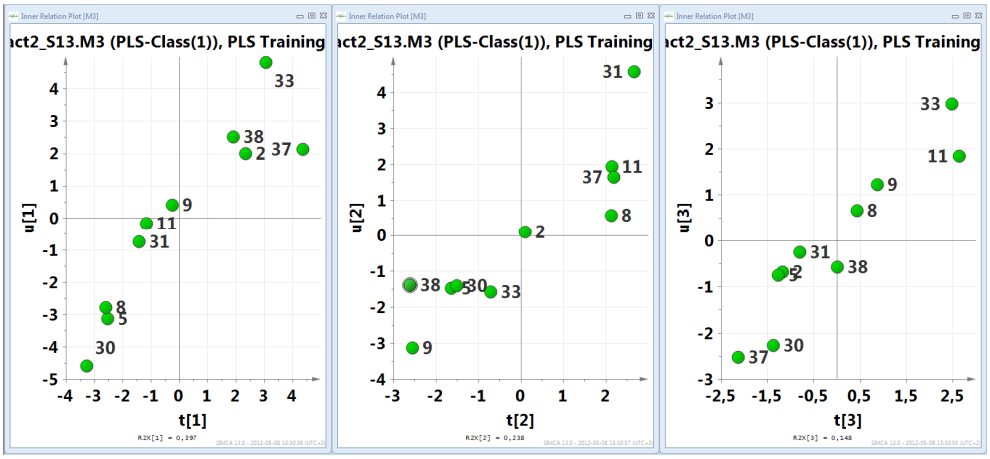


Figure 7.15: (left) PLS t_1/u_1 score plot of the model based on the SURFACTANT training set. There is a weak deviation from linearity in the first component. The first component explains 38% of the response variation. Figure 7.16: (middle) PLS t_2/u_2 score plot. The second component counterbalances some of the non-linearity seen in the first component. It accounts for 33% of the variation in the responses. Figure 7.17: (right) PLS t_3/u_3 score plot. The third component has a linear inner relation, but describes 7% of the response variation only.

The PLS weights of the first two components are shown in Figure 7.18. The third component only accounts for 7% of the response variation and is not plotted here. The four responses are not completely correlated. However, YTemp is moderately correlated with YDet, YDet is partially correlated with YTox, and YTox is partially correlated with YConc. This partial linking of the responses justifies their treatment in one single PLS model.

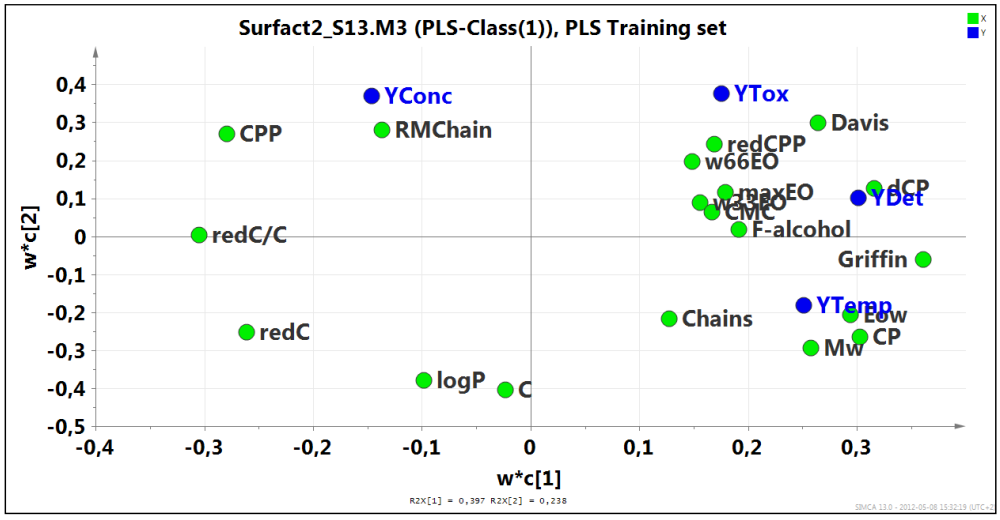


Figure 7.18: PLS weight plot of components 1 and 2. The four responses, highlighted by blue color, are partially inter-related, a fact which justifies using a single PLS model.

We will now give a brief interpretation of the PLS model. A detailed interpretation is found in the original references.

- YDet: In order to achieve high washing efficiency, the surfactant should contain a short carbon chain. This is inferred from the model contribution of the descriptor redC, which denotes the number of carbon atoms in the longest chain of the hydrophobic part.
- YConc: A desired, low value of the surfactant concentration is obtainable for surfactants of a high molecular weight and high molar ratio of dominating chain to total no. of carbons, and low RMChain and CPP.
- YTemp: The washing temperature is influenced by descriptors like molecular weight and the number of EO-units. Low molecular weight and few EO-units correspond to low temperature.
- YTox: Not unexpectedly, the acute toxicity response is well modeled by log P. This kind of relationship is often found in QSAR applications in aquatic toxicology. For low toxicity the surfactants should display low log P (lipophilicity).

Weighting this together indicates that a straight chain –R, with some degree of branching, would provide a good balance between washing performance and toxicity (cf. surfactants 2 and 37).

Additionally, it is interesting to plot, for each response, the relationship between the measured response data and the data predicted by the model. Such plots are shown in Figures 7.19 – 7.22. These figures demonstrate the excellent model performance.

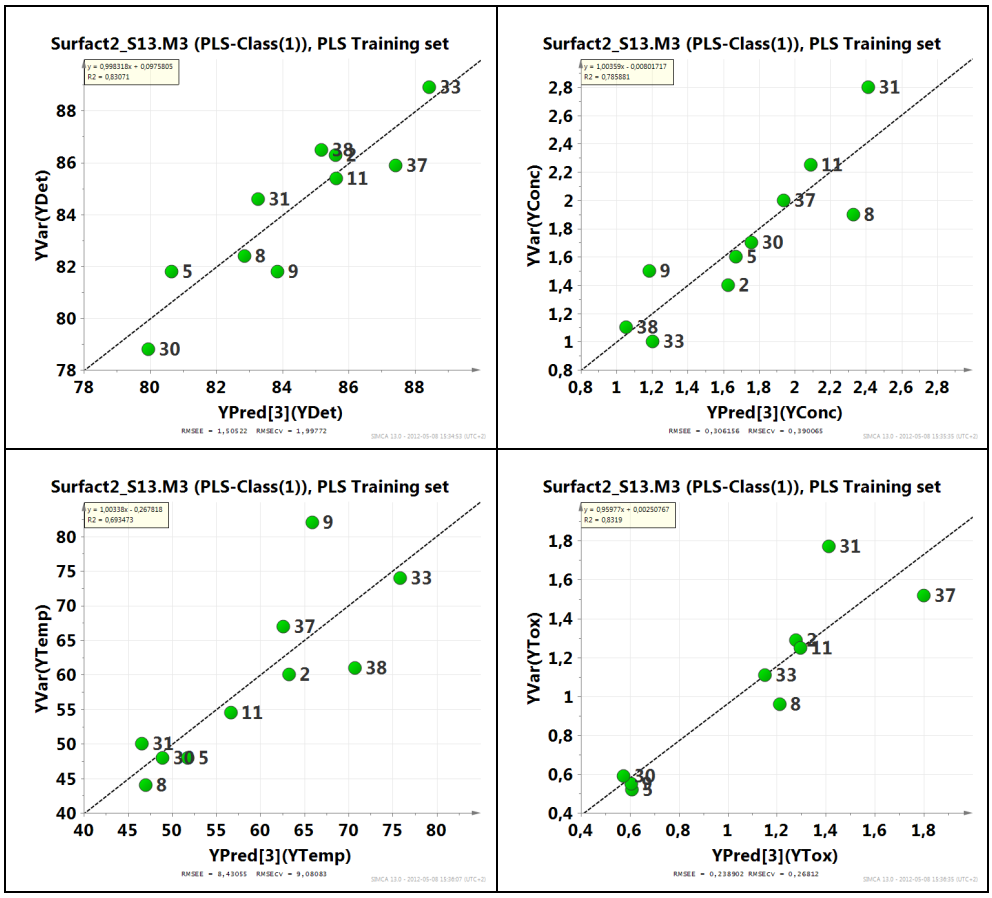


Figure 7.19: (upper left) Relationship between measured data and data predicted by the PLS model for response YDet. Figure 7.20: (upper right) Same as in Figure 7.19, for response YConc. Figure 7.21: (lower left) Same as in Figure 7.19, for response YTemp. Figure 7.22: (lower right) Same as in Figure 7.19, for response YTox.

7.5.6 Predictions of response profiles (Y) for non-tested surfactants

The final step of the SURFACTANT application was to predict Y for the 28 non-tested surfactants. In this process, the known X-data for these compounds were entered into the PLS model, and the unknown Y-data were predicted. There are many graphical ways to display the outcome of such a forecasting procedure. One relevant graph is a plot of the predicted (by the PLS model) score t_1 against the predicted score t_2 for each surfactant included in the prediction phase (Figure 7.23).

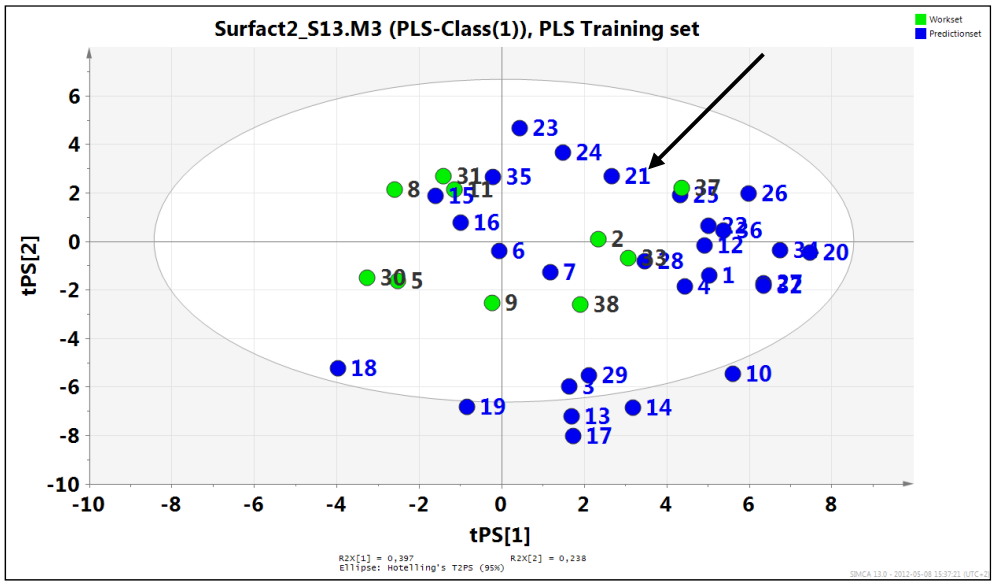


Figure 7.23: Scatter plot of predicted PLS score t_2 versus predicted PLS score t_1. The ten surfactants belonging to the training set are colored green. The arrow marks the location of interesting surfactant molecules.

Figure 7.23 summarizes the predicted scores of all 38 surfactants. The training set surfactants are colored green. The arrow indicates the location of interesting surfactant structures like 21- 26. These are found in the vicinity of surfactants 2 and 37. This sub-set of non-tested surfactants are primary candidates for further exploration of washing performance, toxicity, and biodegradation.

7.5.7 Epilogue

Non-ionic surfactants of the polyethylene type are the subject of intense research and rapid change. Lindgren and coworkers were able to identify a set of promising surfactants by using the methodology described in this chapter. However, the exact identity of the surfactants finally proposed has not been disclosed for proprietary reasons.

The primary conclusions of this application were:

- When all surfactants cannot be tested - multivariate characterization and design is useful in selecting representative compounds.
- Strong relationships exist between measured physico-chemical properties of surfactants and their performance profiles.
- Surfactant performance is a multivariate property and must be addressed by measuring a set of response variables.
- Predictions from the PLS model are useful in identifying interesting surfactants for further performance optimization.
- Systematic measurements of biodegradability are essential, but are lacking.

In concluding this example, it can be said that the approach was so successful that further research was initiated on another set of surfactants, a series of so called sugar-tensides (containing a sugar moiety) [Uppgård, *et al.*, 2000].

7.6 Questions for Chapter 7

1. In which situation is multivariate characterization useful?
2. What are the main steps of multivariate characterization?
3. What is a principal property?
4. What is multivariate design?
5. Which kind of experimental designs are often used in conjunction with design in principal properties?
6. In what way can a set of representative compounds be used for fine-tuning the properties of a class of compounds?

7.7 Summary and discussion

This chapter has given an introduction to multivariate characterization. Multivariate characterization enables a quantification of a discrete problem. This methodology is general and is useful for finding a set of representative observations, such as compounds (solvents, catalysts, reagents) in synthetic organic chemistry, potential drug candidates in drug design, promising pesticide structures in agricultural chemistry, stationary phases in analytical chemistry, or influential batches in a set of discretely varying batches of raw materials.

The concept of principal properties (PPs) is essential to multivariate characterization. Because PPs are orthogonal and few, they are well suited to the construction of experimental designs, like factorial, fractional factorial, D-optimal, or space filling designs. This is called multivariate design in general, or statistical molecular design (SMD) in the pharmaceutical industry. A multivariate design is an important selection tool for finding representative and diverse observations in complicated industrial applications.

Onion design represents a recent addition to the plethora of designs encountered in the SMD toolbox [Eriksson, *et al.*, 2004a; Olsson, *et al.*, 2004a; Olsson, *et al.*, 2004b]. It is a flexible design approach relying on a combination of the best properties of other design families, notably the model support property of D-optimal design and uniform coverage ability of space-filling design.

The onion design splits the candidate set into a number of subsets (“shells“ or “layers”), and a D-optimal selection is made from each shell (Figure 7.24). This makes it possible to select representative sets of molecular structures throughout any property space with reasonable design sizes. The number of selected molecules is easily controlled by varying (i) the number of shells and (ii) the model on which the design is based. The applicability of onion design to a pharmaceutical QSAR problem has been reported [Eriksson, *et al.*, 2004a]. The example dataset contained 967 drug-like molecules. The biological activity under investigation was the inhibition of the major human drug-metabolizing enzyme cytochrome P450 3A4. Onion design was used to select an informative training set. QSAR modeling was thereafter accomplished by means of multivariate projection methodology.

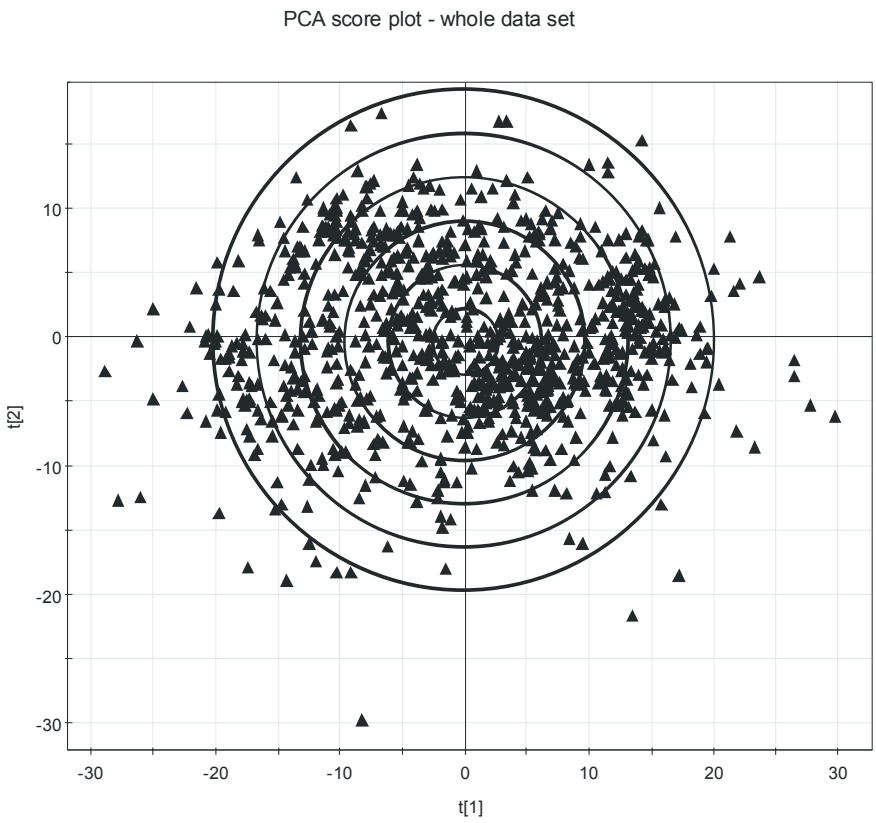


Figure 7.24: Schematic representation of an onion design in a PCA score plot. Prior to setting up the onion design, some parameters have to be defined. The number of principal components, design layers, and selected compounds will influence the shape of the onion design in addition to the type of model supported inside each shell.

8 Multivariate calibration

8.1 Objective

The objective of Chapter 8 is to discuss one of the important areas where PCA, PLS and OPLS have found widespread applicability, namely multivariate calibration. The chapter starts with a discussion of the problems and limitations of traditional calibration, then continues with a description of multivariate calibration. The main steps of multivariate calibration are described. Three examples are used as illustrations: the already familiar SUGAR dataset, and two new datasets called SOIL_DNA and SAWDUST. The SOIL_DNA dataset is unusual in that it contains curve forms based on DNA fingerprinting of heavily polluted soils.

8.2 Introduction

8.2.1 An etymological consideration

According to the Longman Dictionary of Contemporary English [Longman, 1984], the word *calibrate* derives from the process of measuring the inside diameter, the *caliber*, of a tube or a gun. Over the years, this word has come to represent the act of "marking or correcting degrees and dividing points on the scale of a measuring instrument" [Longman, 1984]. And, to quote Harald Martens and Tormod Naes, two chemometricians who have pioneered the development of multivariate calibration, *calibration* is the art of using "empirical data and prior knowledge for determining how to predict unknown quantitative information **Y** from available measurements **X**, via some mathematical transfer function" [Martens and Naes, 1989]. Thus, nowadays, the objective of calibration is often to replace a target or reference method, which may be time-consuming and laborious, with an alternative measurement technique, which should be fast and precise, and preferably also non-destructive to the analyzed samples.

Calibration has found widespread use and acceptance in analytical chemistry and other analytically oriented sciences. This because of the necessity for predicting concentrations of one or several analytes in "unknown" samples, based on information from a set of "known" samples with "known" concentrations of the interesting analytes. For instance, multivariate calibration is increasingly used in process industries for fast, on-line determinations of product quality, such as the octane rating of gasoline, kappa number of pulp, or ash content of white sugar. Multivariate calibration is also a cornerstone element in process analytical technology (PAT), which is further discussed in Chapter 18.

8.2.2 Calibration – What is it?

The traditional approach to calibration is univariate in nature (Figures 8.1 and 8.2). Analytical samples with *known* concentration (C_i) of an analyte are used as the basis for a calibration curve. Each sample is subjected to a measurement technique, for instance HPLC, to establish a signal indicative of the concentration of the analyte. The resulting signal amplitudes (A_i) are then used to construct a calibration relationship, indicated by the line in Figure 8.1 and often called a "standard curve". In the next stage, this relationship is used for new samples with unknown concentrations of the targeted analyte. By characterizing the new sample with the alternative measurement technique, and by entering the resulting signal

amplitude into the established calibration curve, it is possible to estimate the unknown concentration of the analyte (Figure 8.2).

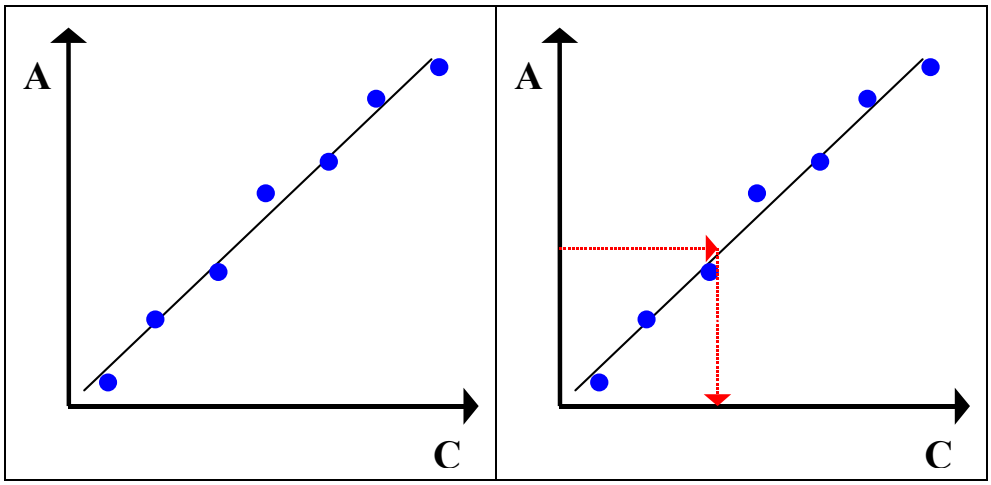


Figure 8.1: (left) Samples with known concentrations (C_i) are analyzed with a measurement technique, such as, HPLC, and the resulting signal amplitudes (A_i) are used in the formation of a calibration graph, a "standard curve". Figure 8.2: (right) At a later stage, the calibration curve is used for estimating the concentration of the investigated analyte in unknown samples. This is accomplished by measuring the signal amplitude of the unknown sample, followed by the estimation of the concentration via the established calibration curve.

In summary, traditional calibration means that known reference concentrations of an analyte are related to data arising from another measuring technique. The relevant calibration relationship is usually established by means of regression analysis. Note, however, that with respect to the regression formalism we are conducting an "inverse" calibration, because C_i is estimated through knowledge of A_i, which is different from the usual manner in regression where "x gives y". After the establishment of a calibration curve, the analyst in most cases checks the linearity of the calibration graph, and often restricts the working range to its linear part.

Traditional calibration normally utilizes only one predictor variable. Usually, however, more than *one* measured x-variable is available, especially when samples have been characterized by spectroscopic or chromatographic methods. In such cases *multivariate calibration* can be used; this involves the use of *all* the information in a battery of X-variables ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$) to model and predict the target analyte concentration ($\mathbf{y}$).

8.2.3 Limitations of traditional calibration

In this section, we will describe some problems with traditional calibration, i.e., (i) selectivity, (ii) precision, and (iii) diagnosis [Wold and Josefson, 2000]. Firstly, in chemical and process analysis it is often difficult to measure signals that are *selective* for just the constituents that are of interest, and, in fact, there is usually no signal where only the target analyte absorbs (Figure 8.3). This is because data may be contaminated by chemical and physical interferences arising from phenomena in the samples themselves, and may also be affected by experimental interferences in the measurement procedure.

In univariate calibration, interfering substances will give accuracy problems, i.e., systematically too high or too low values will be obtained. This accuracy problem is often difficult to detect. A common remedy has been to first run a chromatogram, e.g. by HPLC, to ensure that only one absorbing species is present each time. In the case of multivariate calibration, however, when a new sample has different spectral features, we will recognize this as an outlier in the scores (t) or residuals (DModX). These diagnostics can be used to guard against incorrect inferences.

Secondly, we have the *precision problem*. This arises because noise in a single incoming signal amplitude is transmitted to the predicted concentration of a new sample (Figure 8.4). The advantage of multivariate calibration is that by using more than one signal, i.e., a

multivariate set of signals, it is possible to stabilize the calibration relationship and "filter" the signal so that it becomes "sharper". This is further discussed below.

Thirdly, traditional calibration has a *diagnosis problem*. The calibration or standard curve is valid only for samples similar to the calibration samples. Differences between new samples and the calibration samples may rapidly render the calibration curve unsuitable. This may also manifest itself as a selectivity problem, see (i) above. Furthermore, non-linearities often present an additional challenge in that instruments seldom respond linearly to changes in the constituent concentrations or to changes in the levels of the interferents. Multivariate calibration provides detection tools for outliers and non-linearities, which will aid in the postulation of a more robust and useful calibration model.

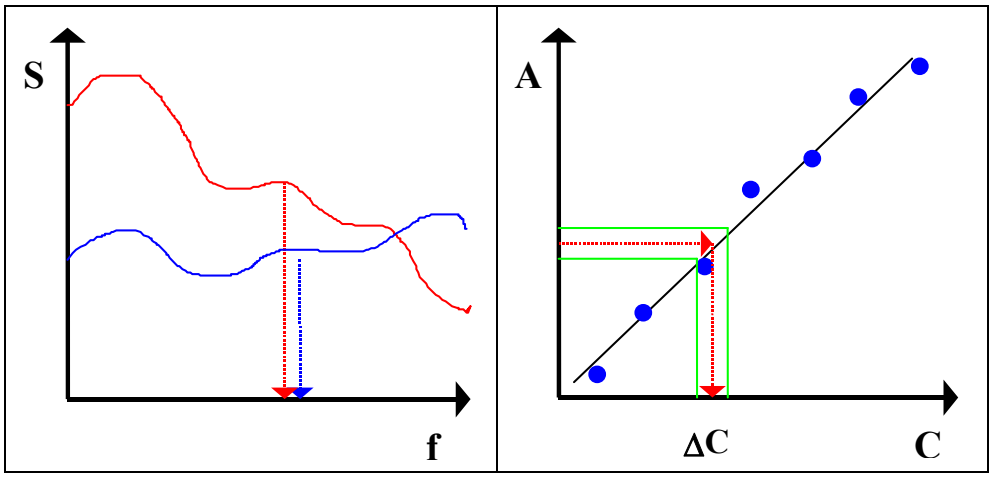


Figure 8.3: (left) Problems with traditional calibration: There is no single signal where only the analyte absorbs. In the diagram above, the capital letter S denotes the signal amplitude, and lower case f the signal frequency. Figure 8.4: (right) Problems with traditional calibration: Noise in signal amplitude transmits to the estimated concentration of a new sample.

8.3 Multivariate calibration

8.3.1 Introduction to multivariate calibration

The problems of traditional calibration related to selectivity, precision, and diagnosis, are less pronounced in multivariate calibration. This is because the latter approach uses the regularities in all data points ("signals") to construct the calibration model, which reduces the noise. In addition, the model residuals are utilized to derive tolerance volumes around the model. These can be used to detect atypical samples that cannot be reliably predicted.

Conceptually, multivariate calibration is closely linked to the philosophy of latent variables and latent variable models. We cannot see such latent variables directly, but only characterize them indirectly through the use of a multitude of measured variables. In calibration this means that we employ not just one selected signal from a spectrum, but rather the entire spectrum is used to encode the "latent signals" in the data. This is illustrated in Figure 8.5.

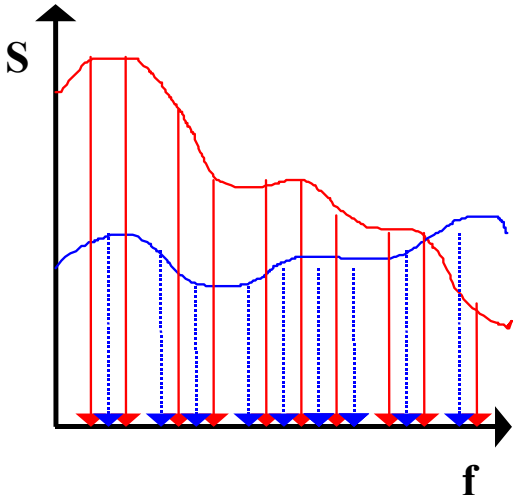


Figure 8.5: In multivariate calibration many signals are used instead of one. Each sample spectrum is digitized at K different wavelengths, which gives K variables ("signals"). In the diagram above, the capital letter S denotes the signal amplitude, and lower case f the signal frequency.

Translating a spectrum to discrete variables is called *digitization*. Each spectrum is "cut" into small equal sections along the wavelength axis (Figure 8.5). (Note, that with modern spectrometers this is really not a practical problem, since measured data are obtained directly in vector form.) If the spectrum is digitized at K wavelengths, K variables (signals) are the result. These data are then analyzed with multivariate projection methods like PLS and OPLS, which can summarize all the signals into a few "latent signals". These "latent signals" are used for calibration against the concentration data. This is beneficial for the modeling and prediction precision, because the "latent signals" are more stable and robust than a single signal (of a single wavelength).

8.3.2 A geometric interpretation of multivariate calibration

In this section we will consider the geometry of multivariate calibration. The dimensionality of the X-space, the "spectral" space, is given by K, the number of digitized wavelengths, which is equivalent to the number of X-variables. Each such variable defines one co-ordinate axis in the K-dimensional X-space. Figure 8.6 serves as an illustration when K = 3. Points, lines, planes, distances, and other geometrical concepts, have similar properties in K-space to those they have in two- or three-dimensional spaces. Each spectrum (each row) in a data table corresponds to a point in the K-dimensional X-space.

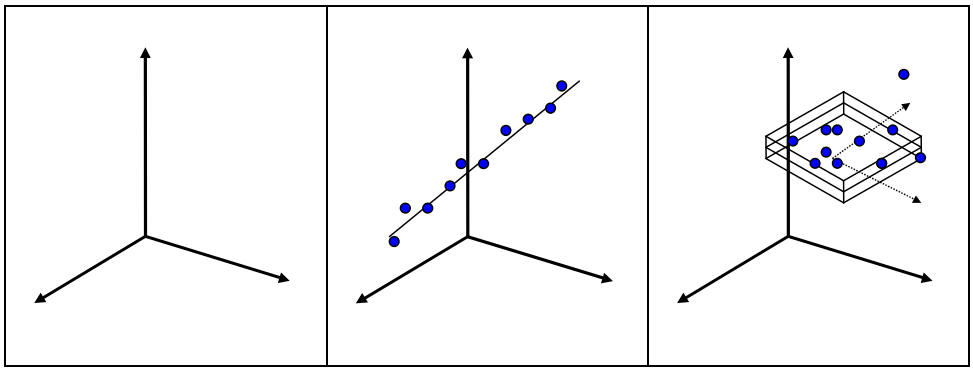


Figure 8.6: (left) When spectra are digitized at K wavelengths, K variables are defined. Each such variable defines one co-ordinate axis in the K-dimensional X-space. Points, lines, planes, distances, etc., have similar properties in K to those in two and three dimensions. Each spectrum corresponds to a point in the K-dimensional X-space. Figure 8.7: (middle) With only one analyte, all points (digitized spectra) are situated on a line, with some deviations caused by noise. Figure 8.8: (right) When one analyte plus some interacting compounds are analyzed, the points are situated on a plane or a hyper-plane, with some deviations caused by noise. The same geometrical interpretation applies when several analytes plus interacting species are treated together.

A data table containing N observations (rows, samples, spectra) corresponds to N points in the K-dimensional X-space. Usually, these points form a swarm with some regularity, whose shape and elongation depends on the complexity of the investigated samples. When the series of investigated samples contains only one analyte, all points are situated on a line, or a smooth curve, with some small deviations caused by noise (Figure 8.7). Further complexity is introduced in the series when the samples also contain some interacting compounds. Then, the points can be represented by a plane or a hyperplane plus some noise (Figure 8.8), with one component (dimension) in the hyperplane per compound (analyte). The latter geometrical interpretation also applies to the case when several analytes and interacting species are analyzed simultaneously.

From the discussions in Chapters 3 and 4, we realize that calibration data may well be analyzed by multivariate projection methods. In fact, the combination of spectroscopic data, such as NIR, and multivariate methodology, such as PLS and OPLS, has proved very successful for multivariate calibration in recent years.

8.3.3 Main steps of multivariate calibration

There are six main steps of multivariate calibration [Wold and Josefsson, 2000]. Five of these pertain to the model training phase, and the sixth is the prediction phase. These steps are:

1. Specification of the analytes with concentration ranges. Selection of the instrumental technique, including consideration of the appropriate wavelength-interval, reflectance or transmission mode, etc.
2. Selection of a representative and diverse set of calibration samples. Ideally, the calibration set should include all variations in X and Y that are relevant to the problem, and which are expected to occur over time. It is particularly important that the calibration set samples span the range of analyte concentrations and also the relevant concentration ranges of interferents.
3. Measurements of X and Y. The multivariate signals, typically spectral or chromatographic data, are recorded for the calibration samples. The analyte concentrations are measured by the reference method.
4. The measured data are investigated for the presence of atypical samples (outliers) and other anomalies. Subsequently, the data are preprocessed and transformed to a form suitable for the subsequent multivariate data analysis.
5. The calibration model is established. This is a crucial step and involves model refinement and interpretation. This step also includes checking for linearity and determination of precision, accuracy, and other measures indicative of model performance. Confidence intervals on model parameters and predicted concentration values are also construed.
6. The last step is the validation and prediction stage. Initially, the calibration model is tested on a relatively small, closed test set. Diagnostics for dissimilarity (outliers) are checked. When the predictive power of the calibration relationship is satisfactory, the model is applied to new samples in the prediction set. In principle, the prediction set is open-ended and constantly grows with time.

Thus, one of the most important aspects of multivariate calibration is the consideration of which samples to use for establishing the calibration relationship (Step 2 above). As in any modeling, it is of crucial importance to collect diverse samples that represent the variation of interest. This can be accomplished by using different approaches, for instance by means of design of experiments (DOE), sampling strategies, and combinations thereof. Different approaches to DOE and sampling in multivariate calibration are discussed in more detail in Section 8.8.

Once a representative and diverse calibration set has been acquired, measurements leading to X and Y are conducted (Figure 8.9). We recall that the Y-data are usually measured using the reference method. Alternatively, the calibration samples can be given “known” analyte

concentrations by using known amounts of analyte and solvent according to a design (see explanation in Section 8.8). For the same set of samples, one also conducts a characterization based on the alternative spectral or chromatographic method. This forms the X-data table (Figure 8.9).

In the next step, the X- and Y-data are related through PLS or OPLS and the calibration relationship is thus established. Finally, the calibration model is employed to classify new samples, and to make concentration predictions for them (Figure 8.10). In this context, it is of course important that the resulting PLS or OPLS model should be valid and free from outliers. Such assessments are facilitated by the use of PLS and OPLS diagnostic tools, as discussed in Chapters 4 and 5 (see also Sections 8.4 and 8.5).

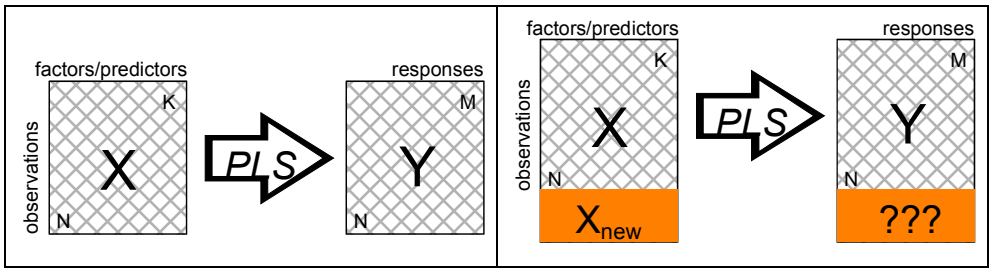


Figure 8.9: (left) An important step in calibration is to collect calibration samples that represent the main variation in the system or process under study. If possible, it is recommended to use design of experiments. For each sample, the Y-data are measured using the conventional analysis method and the X-data by means of the new spectral analysis technique. In order to create the calibration model, the PLS methodology is a good choice. OPLS is another choice. Figure 8.10: (right) When a calibration model has been established, it is used for the classification of new samples, or to make concentration predictions of the analytes.

8.3.4 Some application areas

Multivariate calibration is a growing area, as may be confirmed by browsing the Internet and literature sources. Many interesting applications of multivariate calibration have been published, see e.g. [Halstensen, *et al.*, 2013; Bassbasi, *et al.*, 2013; Tan, *et al.*, 2013; Rodriguez-Pulido, *et al.*, 2013]. We used a common search engine (2012-05-09) with the search phrase "multivariate calibration", and scored more than 1.200.000 entries matching the query. The search was restricted to the years 2000-2012. Table 8.1 summarizes some major application areas, where multivariate calibration has been used as an integral part of research and production.

Table 8.1: Some application areas of multivariate calibration.

• **Wheat, corn, ...**	Protein, water, fat	NIR
• **Peat, coal**	Water, energy, sugars, C, N, S	NIR
• **Lake water**	Humic acids, lignine sulfonate	UV
• **Beer, wine**	Alcohol, protein, sugars, etc.	NIR, IR
• **Whiskey, wines**	Taste, smell, "quality"	GC, HPLC
• **Cellulose, paper**	Raw material, lignine, paper char.	NIR, UV, IR, NMR
• **Pigs (living)**	Fat, protein etc.	NIR
• **PAH**	Environmental monitoring	UV
• **Gasoline**	Octane number	NIR
• **Pharmaceutical**	Drug compounds & metabolites	UV-vis, NMR, NIR
• **Process quality**	Soft sensors	NIR
• + **many, many more**		

8.4 Example: SUGAR

In this section, we outline the SUGAR application in more detail. The SUGAR dataset was kindly contributed by Rasmus Bro at KVL in Copenhagen, Denmark.

8.4.1 Background

At a sugar plant somewhere in Scandinavia, an important quality parameter was the ash content of the sugar, but the conventional measurement technique for this variable was arduous and time-consuming. As a consequence, it was desired to replace traditional wet-chemistry measurements with rapid on-line fluorescence measurements. Multivariate calibration was used to investigate whether this was possible. More details are given in the original literature source [Bro, 1996].

The process data were measured during a campaign of 2.5 days, in the initial start-up period of a new factory. A total of 106 samples (time points, observations) were collected and subjected to fluorescence spectroscopy. Thus, 571 X-variables were created. The spectral data are plotted in Figure 8.11.

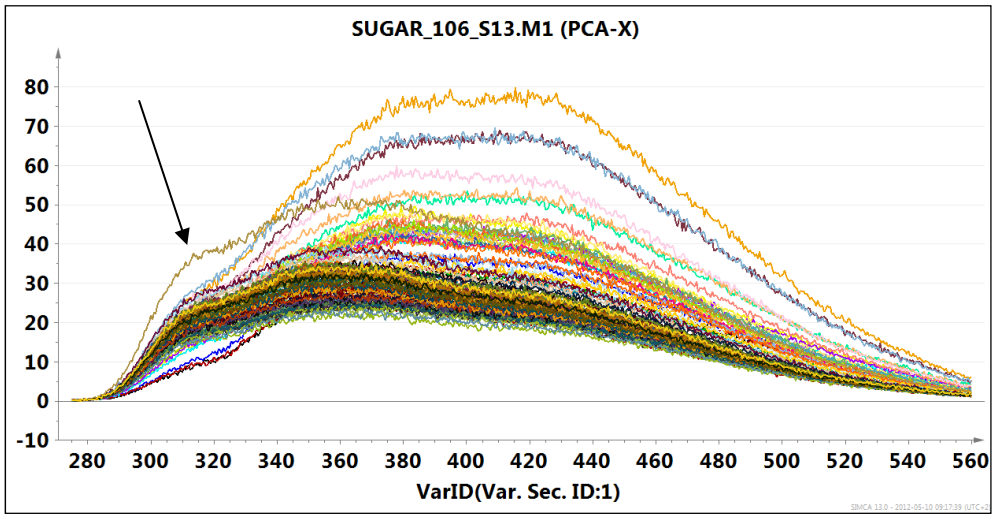


Figure 8.11: Plot of the raw fluorescence data (106 spectra). In the multivariate modeling, each spectrum represents one row of the X-matrix. The arrow indicates an abnormal spectral feature exhibited by sample number 88.

8.4.2 PCA of the spectral data (X)

We here commence the data analysis by computing a PCA model of the X-data, because this may help to detect outliers and other anomalies in the data. However, prior to running PCA, an appropriate means of scaling the X-data was required. Scaling of spectral data is discussed and exemplified in Chapter 13. Since we did not want to blow up noisy variables with low information content, we did not scale to unit variance. Instead, the X-variables were only corrected for their respective means ("mean-centering").

PCA of the mean-centered X-data gave a two-component model, which explained 99% of the variation ($R^2X = 0.99$). The first component is the more important and accounts for more than 94% of the modeled variation. This must be borne in mind when inspecting the t_1/t_2 score plot displayed in Figure 8.12.

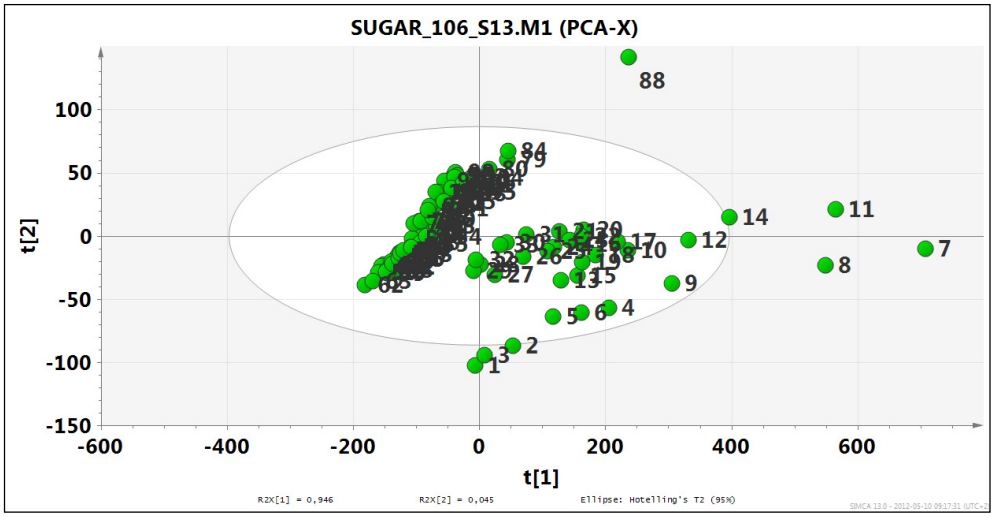


Figure 8.12: PCA score plot of the SUGAR dataset. The two first scores, accounting for more than 99% of the spectral variation, are plotted. The numbering of the observations refers to the sampling sequence. We can see that a marked process drift is manifested by the spectral variables. After process time point 15, a first stage of temporary stability is found, and from time point 35 an area of more representative process operating conditions is established. The ellipse corresponds to the 95% tolerance region, as specified by the T^2 statistic.

In Figure 8.12 we can observe a marked drift at the beginning of the sampling period. Recall that this dataset contains measurements from the initial start-up phase of a new sugar plant. Hence, this movement is not surprising. The score plot suggests that a state of comparatively stable process conditions is achieved around observation 15, and that a really stable region is encountered after time point 35. Also, it is evident that observation number 88 displays deviating spectral behavior. This is related to a "knee" in the spectrum, as indicated by the arrow in the left-hand part of Figure 8.11.

To interpret the meaning of the PCA model we examine the loadings, here the loading vector of the first component denoted p_1 and the loading vector of the second model dimension denoted p_2. Usually, these vectors are plotted together in a 2D-scatter plot, but in the analysis of spectroscopic data the plotting of each individual loading vector against the variable (wavelength) number is more prevalent. Figures 8.13 and 8.14 provide such loading spectra. From these plots we can infer two mathematically independent spectral contributions in the data.

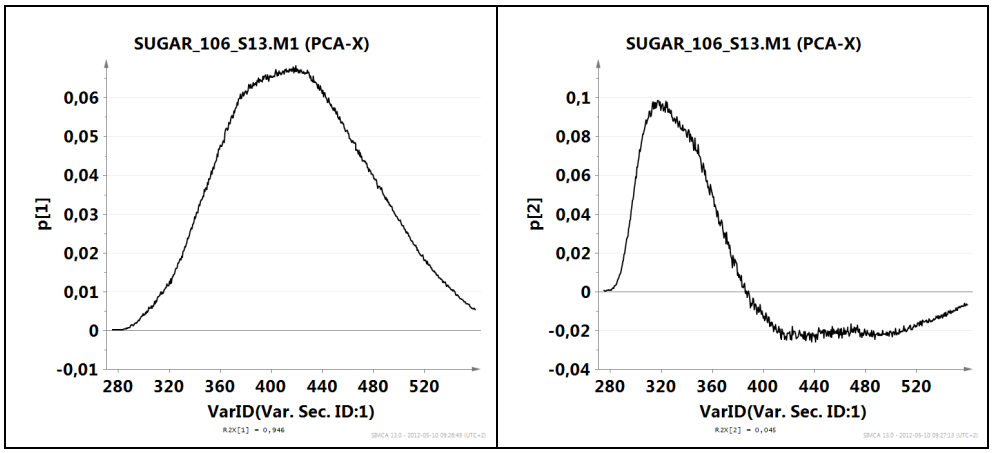


Figure 8.13: (left) PCA loading spectrum of the first component of the PCA model for the SUGAR data, which mimics the average fluorescence spectrum (cf. Figure 8.11). Figure 8.14: (right) PCA loading spectrum of the second component indicating a secondary peak-like structure around variables 50 – 150 (in the 300 – 350 nm region).

The line plot of p_1 (Figure 8.13) resembles the average fluorescence spectrum across the whole set of 106 observations. Also the second loading vector (p_2) reflects a peak-like structure in its left-hand part (at low wavelengths). From the shapes of the raw-spectra, it appears that the little "knee" displayed by some samples at low wavelengths to the left of the main peak might be the phenomenon captured by the second component.

In summary, the PCA model yielded important insights regarding the "movements" of the sugar process plant. The first start-up period seems to last until time point 35.

8.4.3 PLS modeling of the entire dataset

We mentioned in Section 8.3.3 that care must be exercised in the formation of pertinent training and validation sets for calibration. Ideally, the spectral variation in X and the numerical range in Y should be spanned approximately uniformly by such sub-sets of data. In order to accomplish a reasonable division of the SUGAR samples into training and validation sets, we proceeded as follows. The 106 observations were sorted according to a descending ash content (yielding a new dataset called SUGARSRT where all samples are re-numbered according to decreasing ash content). Then the samples were divided in two parts, by allocating every odd-numbered (according to the new numbering system) observation to one group and every even-numbered (according to the new numbering system) observation to the other group.

A beneficial consequence of this sorting and sub-division procedure is that the extent of auto-correlation among the samples is suppressed. This makes the estimation of the goodness of prediction (Q^2Y_{int}) by cross-validation more reliable. Note, however, that this approach, while optimal for this particular calibration modeling, is not necessarily appropriate for all long-term process monitoring.

Two preliminary PLS models were then established, one for each sub-group. By inspecting the inner relations (u_1 vs. t_1) of each PLS model, it is evident that some observations sampled at the beginning of the campaign are not modeled well (see Figures 8.15 and 8.16). In Figure 8.15 observation no. 1 is extreme and in Figure 8.16 there is a considerable scatter around the ideal diagonal in the high ash content region. These findings suggest that some of the initially sampled observations might have to be omitted from the calibration modeling, since they represent chemically irrelevant production conditions.

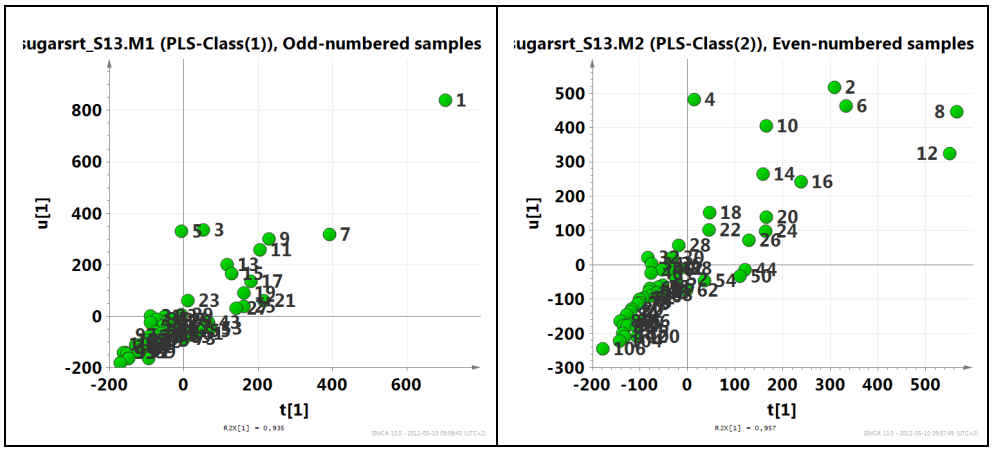


Figure 8.15: (left) PLS t_1/u_1 score plot of the odd-numbered training set, showing the extreme behavior of sample 1, and some minor deviations off the ideal diagonal by samples 3 and 5. Note that the numbering system in this plot refers to the ash content sorted data, which is different from the sequential time order numbering used in Figure 8.12. The maverick sample 88 of Figure 8.12 corresponds to sample 9 in the current plot. Figure 8.16: (right) PLS t_1/u_1 score plot of the even-numbered training set, displaying considerable scatter of observations in the high ash content region. The numbering system is the same as in Figure 8.15.

The question then arises: how many observations should be eliminated? A first glance at the score plot of the PCA (Figure 8.12) might suggest that the first 35 samples plus the deviating observation number 88 should be skipped. However, a simultaneous inspection of

Figures 8.12, 8.15 and 8.16 indicates that after time point 15 the sugar process enters a temporary stage of mild shifts in operating conditions. This region is also rather close to the main cluster seen in Figure 8.12. Based on these PCA and PLS findings, we decided to skip 16 samples, the first 15 plus number 88.

8.4.4 PLS modeling of the reduced dataset

In the reduced SUGARSRT data set (with 90 observations), the sorting of the data according to ash content, and the division into odd- and even-numbered groups of observations, were all preserved. This gave odd- and even-numbered sub-sets comprising 45 observations each.

The fitting of PLS models to these two sub-sets produced two-dimensional models of sound explanatory and predictive powers (Figures 8.17 and 8.18). For the odd-numbered training set, the model obtained had a goodness-of-fit of 81% ($R^2Y = 0.81$) and a goodness-of-prediction of 79% ($Q^2Y_{int} = 0.79$). The corresponding values for the model based on the even-numbered training set were 84% ($R^2Y = 0.84$) and 82% ($Q^2Y_{int} = 0.82$). The subscript "int" of the Q^2's denotes that these estimates were derived by cross-validation. The likeness of these R^2- and Q^2-values indicates that the two sub-sets are rather well-balanced and have similar characteristics. Hence, the sub-division of observations appears to have been warranted.

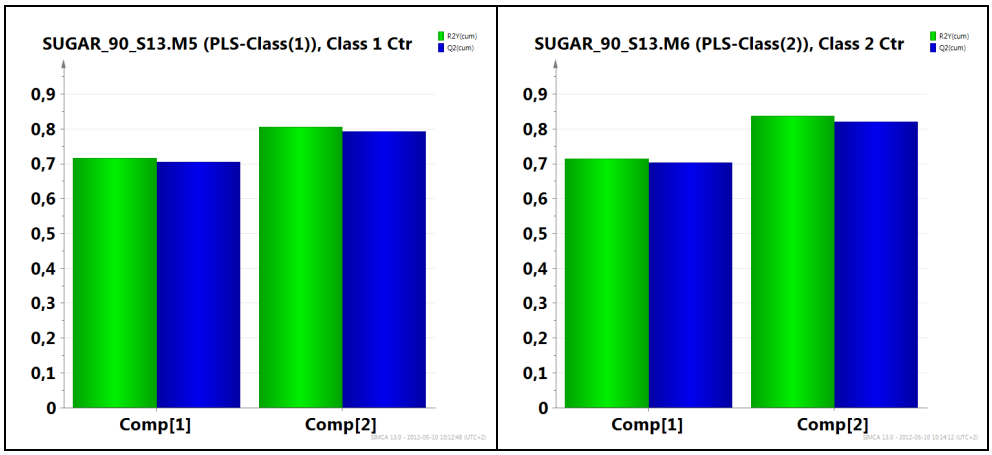


Figure 8.17: (left) Model summary of the PLS model for the odd-numbered training set. After two components $R^2Y = 0.81$ and $Q^2Y_{int} = 0.79$. Figure 8.18: (right) Model summary of the PLS model for the even-numbered training set. After two components $R^2Y = 0.84$ and $Q^2Y_{int} = 0.82$.

By using the two sub-sets as external validation sets, it is possible to further test the predictive power of the two models. We obtained the results summarized in Figures 8.19 and 8.20. The external Q^2-values, Q^2Y_{ext}, amount to 0.83 and 0.80, which are almost identical to the internal Q^2-values of 0.79 and 0.82. Based on these model performance statistics, we conclude that both cross-validation and validation by external prediction yield similar estimates of the predictive power.

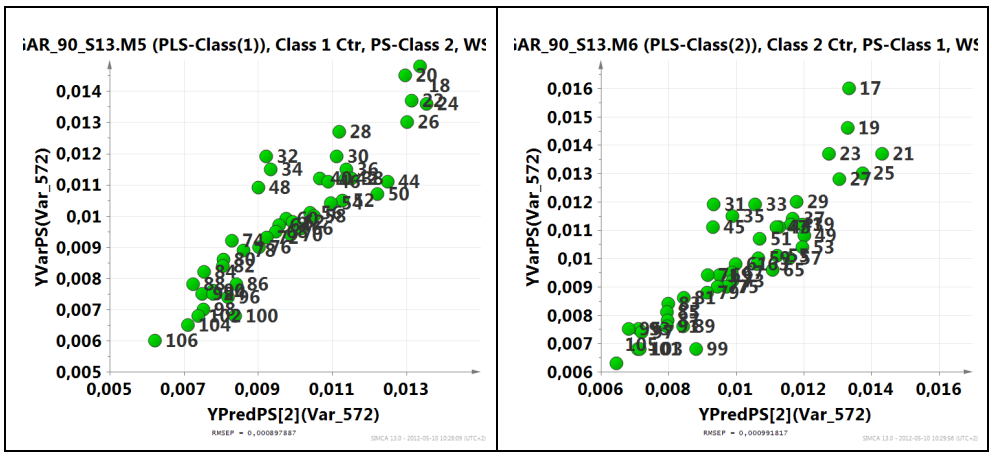


Figure 8.19: (left) Relationship between observed and predicted ash content for the odd-numbered training set/even-numbered validation set. Q^2Y_{ext} amounts to 0.83. Numbering as in Figure 8.16. Figure 8.20: (right) Relationship between observed and predicted ash content for the even-numbered training set/odd-numbered validation set. Q^2Y_{ext} is equal to 0.80. Numbering as in Figure 8.15.

In multivariate calibration it is sometimes possible to enhance the predictive power by means of signal correction. Signal correction as a concept is addressed in Chapter 14.

8.4.5 OPLS modeling of the reduced dataset

In order to deepen the understanding of the relationship between the fluorescence data and the ash content, two parallel OPLS models were fitted to the two sub-sets. Both OPLS models were 1 + 1 + 0 models, which agree with the dimensionality of the two previous PLS models (containing two components each). Hence, there is no difference between each pair of PLS and OPLS models regarding external predictive power. The interesting aspect is that OPLS uncovers how much of the variation in the fluorescence data is predictive (correlated) and orthogonal (uncorrelated) to the ash content. As shown by Figure 8.21, for the odd-numbered sub-set 81.4% of the spectral variation is predictive to the ash content. However, as much as 17.2% of the spectral variation is orthogonal to the ash content. The corresponding numbers for the even-numbered sub-set are 75.6% predictive and 22.7% orthogonal variation (no plot shown).

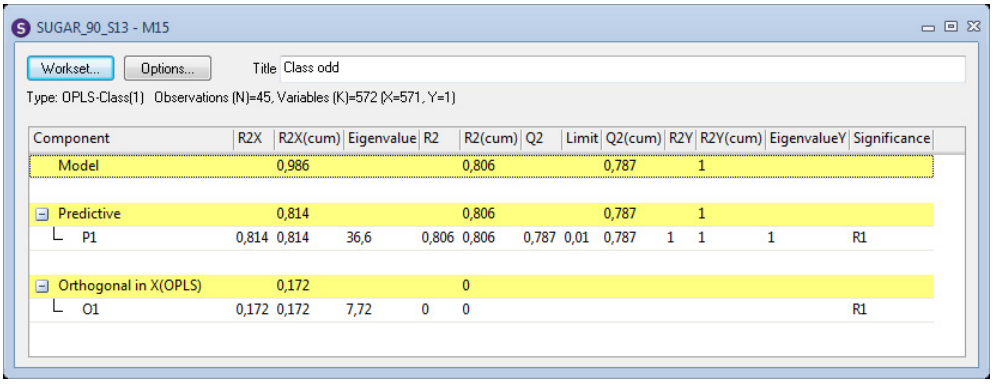


Figure 8.21: Model Window for the OPLS model computed for the odd-numbered training set. It shows that 81.4% of the variation in the fluorescence data is predictive to the ash content. As much as 17.2% of the spectral variation is orthogonal to the ash content.

8.4.6 Summary of SUGAR application

The conclusion of the SUGAR application was that it is possible to use fluorescence measurements for modeling and predicting the quality of produced sugar. Predictive abilities (Q^2) in the range 0.80 - 0.85 must be considered as very satisfactory considering the on-line application. Such results have also been obtained in additional studies performed at other sugar plants across Scandinavia, where models of similar or even higher predictive abilities have been established.

Furthermore, the delineated example is good in that it demonstrates the purposeful sequence of first conducting PCA, then following it up with PLS and OPLS. With PCA it is possible to get a good understanding of what is going on in a dataset. In this way, the analyst is better prepared for the ensuing PLS and OPLS analyses.

8.5 A second example: SOIL_DNA

8.5.1 Background

In many cases, multivariate calibration is practically the same as PLS modeling of spectroscopic data. Multivariate calibration may also be applied to other types of curve forms which are not spectroscopic or chromatographic in nature. Examples of such curves might be particle size distributions, tear index curves of paper, curves for studying rheology of complex products, and so on.

The current example, which we call SOIL_DNA, originates from soil remediation studies. Soil remediation is, among other things, concerned with clean-up and recuperation strategies for soil contaminated by oil spills or other undesired exposures. In polluted soils, organic compounds are metabolized and mineralized as a result of microbial action. Thus, it is conceivable that a relationship exists between exposure profiles of pollutants and the composition of the microbial community. Accordingly, in soil remediation it is of interest to monitor how levels of organic pollutants vary with time after exposure, and to understand how bacterial populations change when exposed to external stress factors.

Per Wikström, at the Swedish National Defense Research Establishment, has approached soil remediation from a multivariate point of view, and has kindly contributed the SOIL_DNA dataset. This dataset has 21 observations, 1018 X-variables and ten Y-variables. Each observation corresponds to one day in October 1998, when the experiments were conducted.

Ten organic substances (ethylbenzene, xylene, decalin, naphthalene, tridecane, heptamethyl-nonane, tetradecane, phenol, p-cresol, and pristane) were added to a tank of artificial soil. For each day during an almost three week long investigation period, the decay curves (concentrations) of these 10 compounds were measured (toluene extraction and gas chromatography). These concentration readings form the ten responses.

Moreover, in order to get a picture of the changing bacterial population, DNA fingerprinting was used to characterize microbial biomass at each sampling occasion. DNA fingerprinting means that PCR (polymerase chain reaction, NOT principal component regression) is used to amplify DNA fragments. During PCR these fragments become labeled with a fluorescent moiety, so that after gel-separation (polyacrylamide gel) they can be scanned by a laser, to establish a “DNA-fingerprint-spectrum”. The DNA fingerprinting information is encoded in the 1018 X-variables.

Figure 8.22 shows a line plot of the 21 raw DNA fingerprints. One fingerprint was acquired for each day of the measurement campaign. More information is available in the literature [Wikström, *et al.*, 1999; Wikström 2001]. In the data analysis below, these fingerprint data were log-transformed.

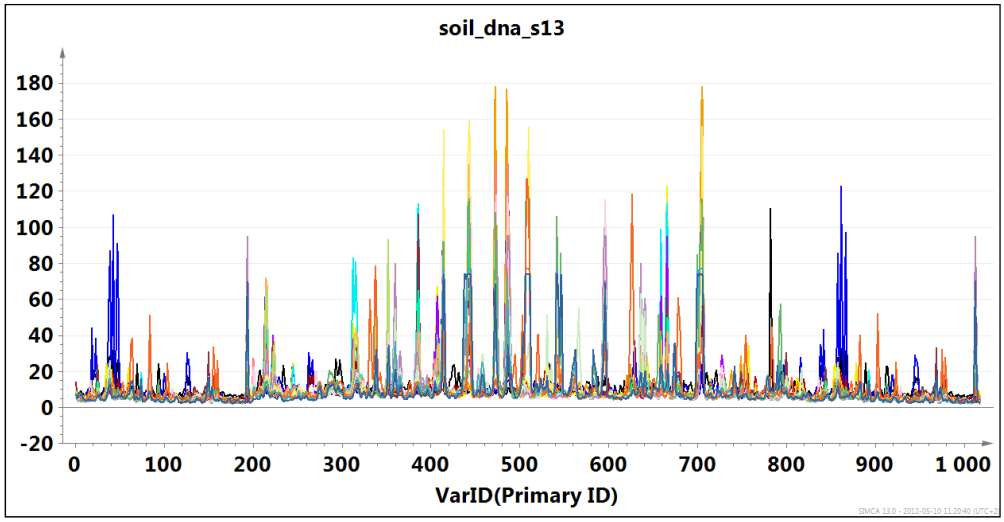


Figure 8.22: Line plot of the raw DNA fingerprint data, where 21 "spectra" are superimposed on top of one another. The y-axis displays signal amplitudes indicative of the amount of each fragment type, while the x-axis represents the different fragment sizes. The DNA fingerprinting technique used enables fragments in the range of 100 to 1000 DNA bases to be seen. The largest fragments are found to the left.

8.5.2 PCA of the fingerprint data (X)

In the first stage of the data analysis, PCA was applied to Pareto-scaled fingerprint data (X). This was done in order to examine whether the data contained any outliers, and to survey possible time trends. According to cross-validation, four principal components were optimal for the description of the 1018 X-variables (Figure 8.23). After four PCs $R^2X = 0.77$ and $Q^2X = 0.53$. Figure 8.24 displays the first score plot, which accounts for 58% of the total variation.

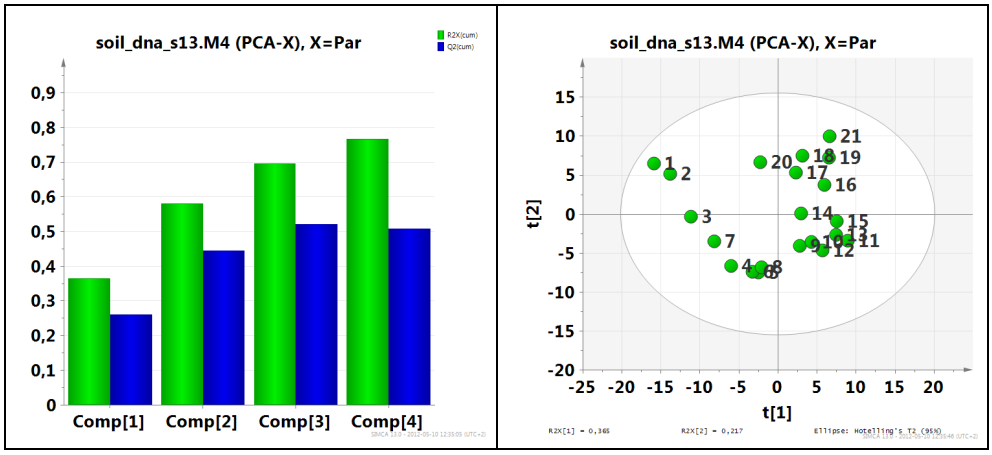


Figure 8.23: (left) Model overview of the PCA of the X-data. According to cross-validation, four components are optimal for the description of the X-data. With four PCs, the explained variation is 77% and the predicted variation 53%. Considering the large number of X-variables, 1018, these values are good and indicate a lot of systematic variation in the X-data. Figure 8.24: (right) PCA score plot with t_2 plotted against t_1. This plot represents 58% of the fingerprint variation.

As shown by Figure 8.24, there is a marked variation over time expressed in the fingerprint data. Remember that each point in this score plot represents one day of the explored time period. The changes are most dramatic at the beginning of the sampling period (low observation numbers) but calm down towards the end of the study. It is encouraging to see that there are no distinct outliers. This indicates that there are no sudden, abrupt, and discontinuous changes in the composition of the microbial population.

In summary, the PCA of the X-data indicates strong time dependencies among the observations, and that there are no alarming outliers.

8.5.3 PCA of the chemical concentration data (Y)

Since the Y-matrix contains ten responses it is recommended to analyze these with PCA before turning to PLS and calibration modeling. The PCA modeling of UV-scaled Y-data resulted in one component with $R^2X = 0.92$ and $Q^2X = 0.88$ (Figure 8.25). However, we decided to compute two components to enable construction of scatter plots. The high R^2X- and Q^2X-numbers indicate that the degradation patterns of the ten organic pollutants are strongly correlated. The R^2X- and Q^2X-values – after the first component – for the individual variables are visualized in Figure 8.26. Evidently, four variables, naphthalene, tridecane, heptamethylnonane, and tetradecane, are extremely well modeled. On the other hand, decalin is not so well explained by the model. This might indicate that there are deviating measurements in this variable.

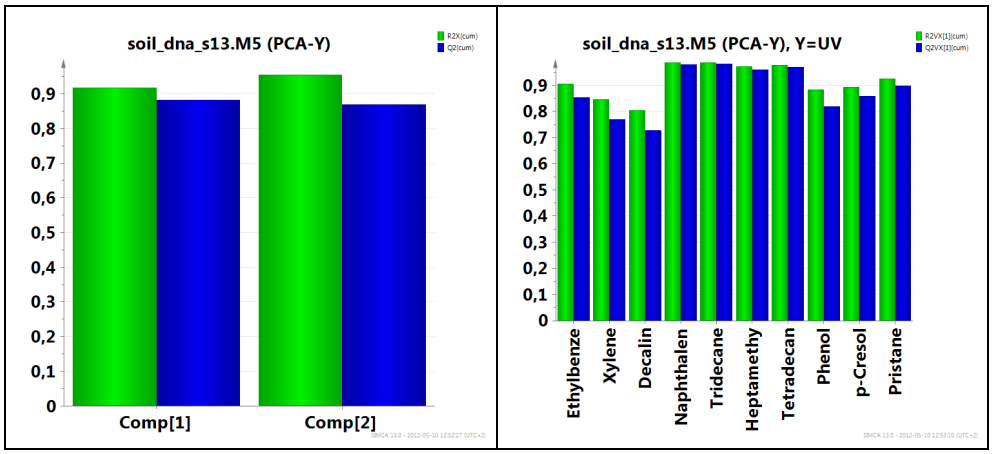


Figure 8.25: (left) Model overview of the PCA of the Y-data. According to cross-validation only the first PC is really important with $R^2X = 0.92$ and $Q^2X = 0.88$. Figure 8.26: (right) Model performance for the individual responses, after the first component. Apparently, four variables, naphthalene, tridecane, heptamethylnonane, and tetradecane, are extremely well modeled. On the other hand, there are variables like decalin for which some improvements would be desirable.

In order to interpret the PCA model, we consider the scores and loadings seen in Figures 8.27 and 8.28.

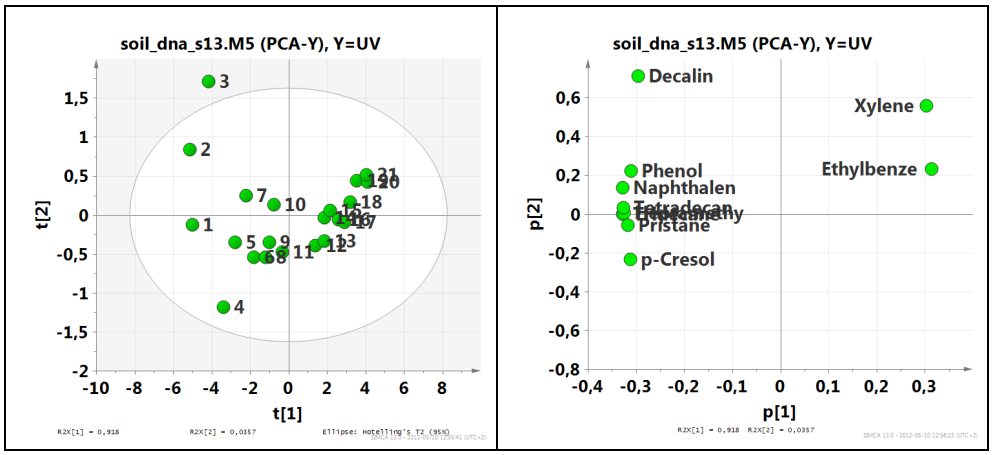


Figure 8.27: (left) PCA score plot with t_2 plotted against t_1. This plot represents 95% of the variation in chemical concentrations. The first component alone accounts for 92% of the variation. Figure 8.28: (right) PCA loading plot with p_2 plotted against p_1. Since the first component dominates the model, we realize that the ten concentration variables are heavily correlated.

As evidenced by Figure 8.27 there is a strong variation over time among the 21 samples. The early samples are found to the left and the later samples to the right. We must remember that the second component is not significant and only represents 3% of the

variation. Hence, the significance of the separation among samples 1, 2, 3, and 4 should not be exaggerated.

In the loading plot, we find, somewhat surprisingly, that eight variables are grouped to the left and two to the right. This means that only for the eight species to the left are the concentrations decreasing over time. The amounts of xylene and ethylbenzene actually increase over time. This might indicate that, as the other eight compounds are metabolized by microbial action, xylene and ethylbenzene are formed.

In summary, the most important finding of the PCA of the Y-data is that the ten responses are highly correlated. As a consequence, in the multivariate calibration of these data, only one PLS model should be required (see Section 8.5.4). There are no strong outliers in the concentration data.

8.5.4 PLS modeling

In the multivariate calibration, the X-block (1018 fingerprint variables) was Pareto scaled and the Y-block (10 concentration responses) was scaled to unit variance. The PLS modeling yielded a three-component model with $R^2X = 0.68$, $R^2Y = 0.91$ and $Q^2Y = 0.80$ (Figure 8.29), which are excellent modeling results considering that ten responses are modeled at the same time. The individual R^2Y- and Q^2Y-values of the responses are shown in Figure 8.30. The majority of the responses are well modeled and well predicted. This means that the fingerprint data contain much information of relevance for explaining the behavior of the responses.

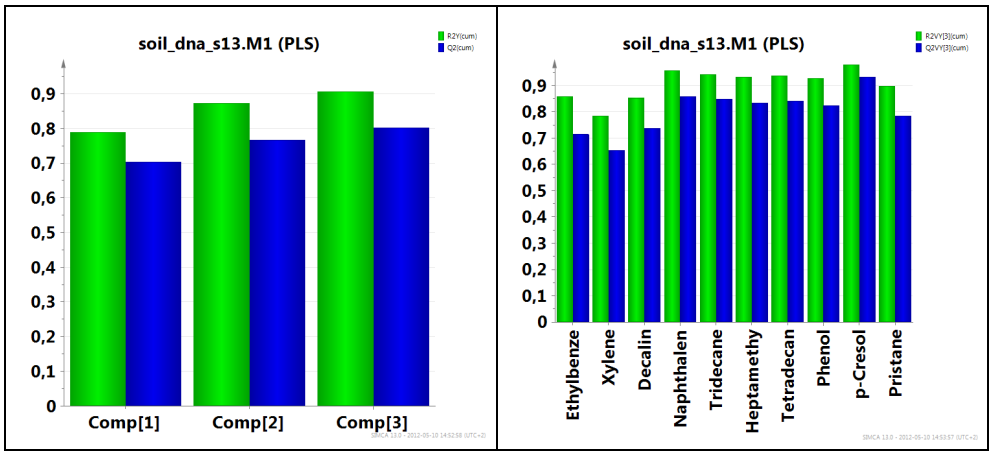


Figure 8.29: (left) Model overview of the calibration model. After three PLS components $R^2X = 0.68$, $R^2Y = 0.91$ and $Q^2Y = 0.80$. Figure 8.30: (right) Individual R^2Y- and Q^2Y-values of the ten responses. It is interesting to note that xylene and ethylbenzene, the two compounds for which the levels rise over time, are the two "worst" modeled responses (they are still well modeled and well predicted!).

Since the overall model statistics point to a good model, we proceed to model interpretation. The three PLS t/u score plots, shown in Figures 8.31 – 8.33, indicate that the X- and Y-data are strongly inter-related. We can observe that samples 11 and 20 display some minor deviation in the first score plot (Figure 8.31). These deviations are corrected for in the second and third components (Figures 8.32 and 8.33).

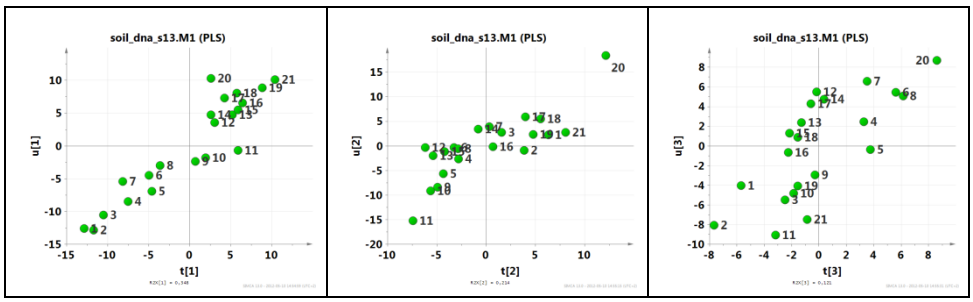


Figure 8.31: (left) PLS t_1/u_1 score plot. The first component explains 79% of the variation in the responses. Each observation corresponds to one day. Obviously, observations 11 and 20 deviate weakly from the main diagonal. However, this non-conformity is modeled by the second and third components. Figure 8.32: (middle) PLS t_2/u_2 score plot. The second component explains 8% of the variation in the responses. We can see that samples 11 and 20 are found at the ends of the correlation band. Hence, we understand that the second component is actually a compensation for some small deviations, which were left unmodeled after the first model dimension. Figure 8.33: (right). PLS t_3/u_3 score plot. The third component only explains 4% of the Y-variation, and therefore it is understandable that the correlation band is weakest and widest for this model dimension.

Figures 8.34 and 8.35 provide weight and VIP plots. In this case, it must be admitted that the model interpretation is somewhat complicated. The rugged appearance of these plots arises from the complex profiles of the raw fingerprint data (cf. Figure 8.22).

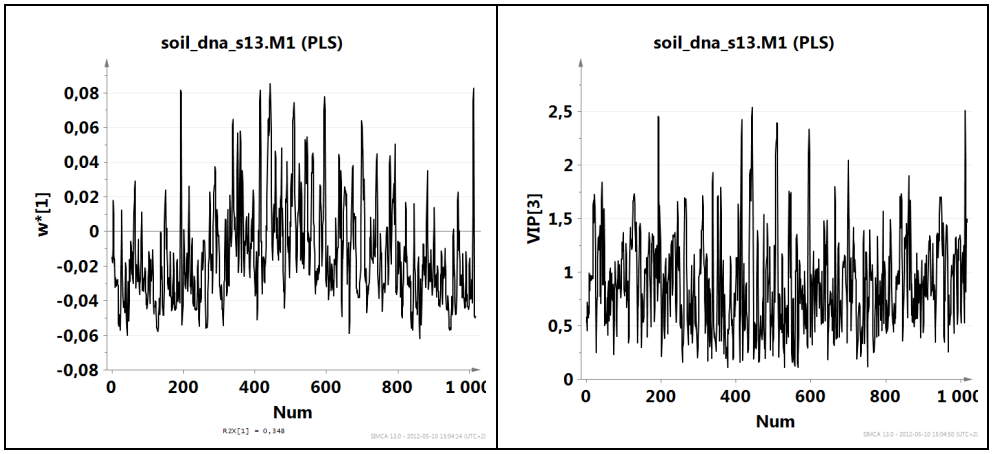


Figure 8.34: (left) PLS weight spectrum of the first model dimension. Figure 8.35: (right) VIP plot after three PLS components. This plot indicates that informative DNA fragments are found throughout the entire fingerprint spectral domain.

Another way of examining the efficiency of the fingerprint data for modeling the response data is to plot the observed response data against the corresponding values estimated by the PLS model. Such plots are shown for two responses in Figures 8.36 (decalin) and 8.37 (p-cresol). Decalin represents a case where there are some discrepancies between observed and estimated response values, whereas the relationship for p-cresol is almost ideal.

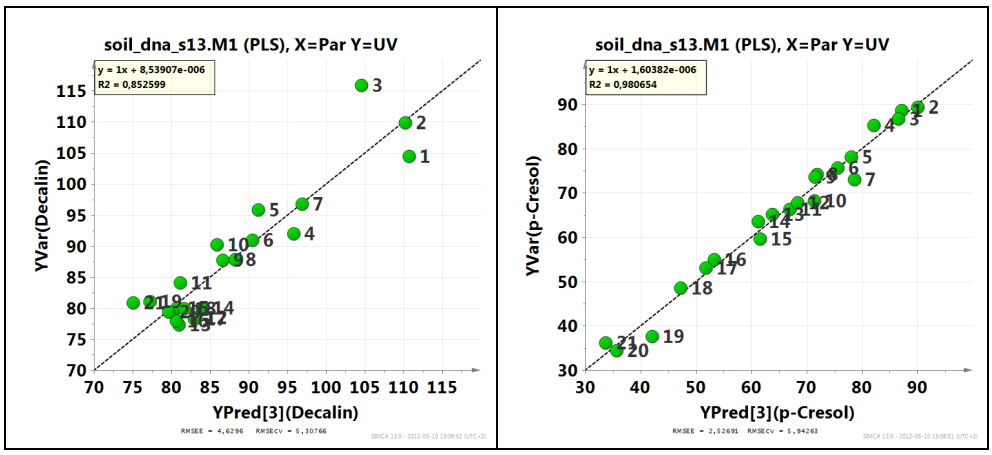


Figure 8.36: (left) Relationship between observed and estimated ("predicted") decalin concentrations. Each plot mark corresponds to one observation. Figure 8.37: (right) Agreement between observed and estimated ("predicted") p-cresol concentrations.

8.5.5 OPLS modeling

In the next step, OPLS was used to analyze the relationship between the X-block and the Y-block. The OPLS model obtained was a 3+0+0 model, i.e., a model consisting only of three predictive components and thus being very similar to the PLS model. The order and size of the three predictive components match the order and size of the three components of the PLS model. The conclusion is that the DNA fingerprint data do not contain any systematic variation that is orthogonal to the concentration profiles.

8.5.6 Summary of SOIL_DNA application

The conclusion of the SOIL_DNA investigation was that DNA fingerprinting is a useful method for mapping the properties of a changing microbial community. This kind of change is much influenced by external stress factors, such as high levels of organic pollutants. Therefore, there exists a close relationship between the fingerprint information and the chemical concentration data. This association of data may beneficially be exploited with multivariate calibration.

It was quite unexpected that the concentration would increase for two of the contaminants, xylene and ethylbenzene. This might be because these compounds are formed as a result of microbial digestion of the other eight substances. Furthermore, there may be other factors, abiotic in nature, which may influence the degradation pathways of these compounds: factors such as binding, solvation or adsorption.

8.6 A third example: SAWDUST

8.6.1 Background

NIR spectroscopy and PLS modeling is a powerful combination for multivariate calibration in the paper, pulp, and forestry industry. In a tutorial application conducted at our laboratory, with the aim of demonstrating the utility of NIR for monitoring wood fiber properties, sawdust from industrial sawing of birch, pine, and spruce were mixed according to a three-factor mixture design. It was decided to tailor the mixture design to support a quadratic regression model. The training set was designed to comprise 15 samples, and the prediction set to contain 12 samples, i.e., a total set of 27 samples was produced. Observe that the prediction set samples were deliberately made slightly different from the training set samples to allow a realistic assessment of the predictive power of the calibration model to be developed.

For each sample, NIR data in the range 1100 – 2300 nm were recorded twice, producing an X-data table with 54 (2 times 27) rows and 1201 variables (Figure 8.38). The corresponding Y-data table was made up of the proportions of birch, pine, and spruce in each sawdust blend. For comparative purposes, Figure 8.39 provides a plot of NIR spectra for pure-component "mixtures", i.e., only birch sawdust, only spruce sawdust, and only pine sawdust. There is no doubt that there are spectral differences between the three types of wood material.

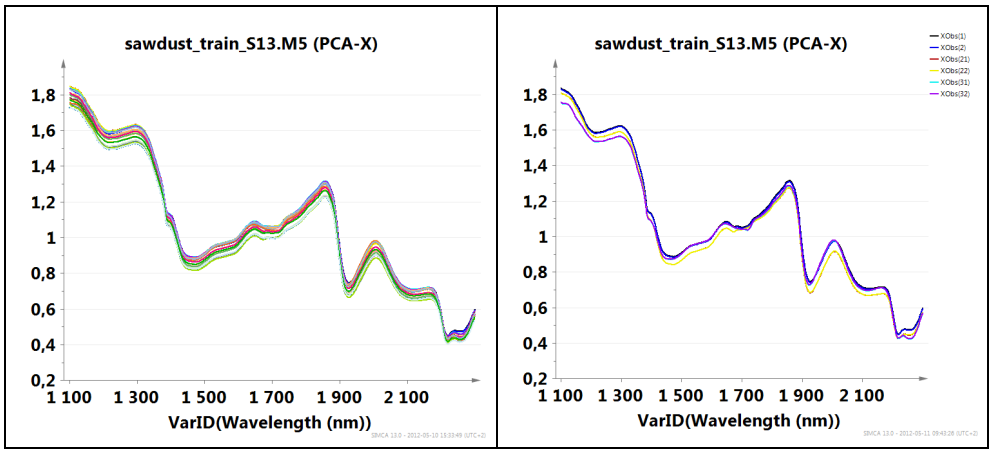


Figure 8.38: (left) Plot of raw SAWDUST data. Figure 8.39: (right) Plot of NIR spectra for pure component mixtures of birch, pine, and spruce sawdust.

8.6.2 PCA of NIR data (X)

The PCA of the X-data of the 30 (2 times 15) training set samples (sawdust blends) resulted in four components explaining 99% of the spectral variation. Figure 8.40 shows the score plot of the first two components, which reflect more than 97% of the total spectral variation. The spread among the samples is apparent, and is due to the underlying mixture design. In order to check the similarity between the training set and the prediction set, the model derived was utilized for classification of the 24 (2 times 12) prediction set samples. The result of this classification procedure is summarized in Figure 8.41. As shown by Figure 8.41, the prediction set samples are somewhat different, but not very different, from the training set samples. Hence, the use of the prediction set for model validation appears reasonable.

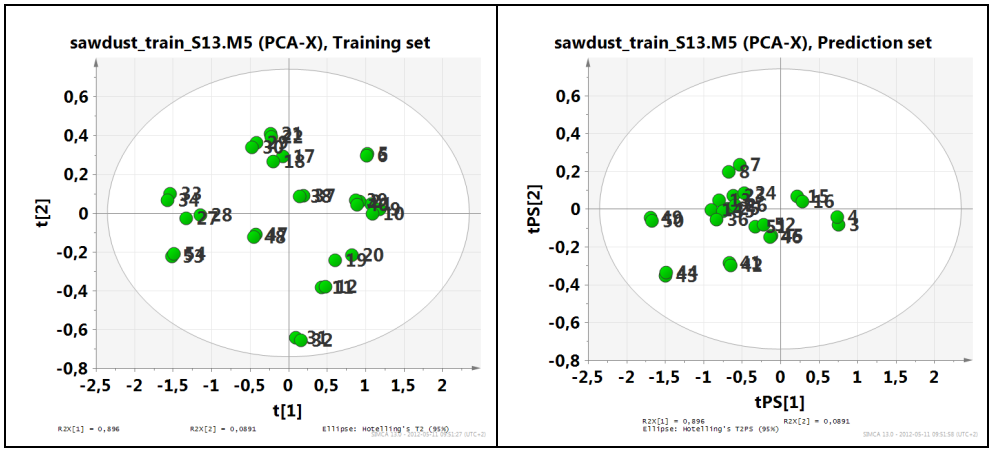


Figure 8.40: (left) PCA t_1/t_2 score plot of the model for the training set of the SAWDUST application. Because two spectra were recorded for each blend of sawdust, pairs of points are seen in the score plot. Figure 8.41: (right) Same plot as in foregoing figure, but the prediction samples have been projected onto the model plane.

8.6.3 PLS modeling

The PLS modeling, using the training set and mean-centered X-data, yielded a model with R^2Y and Q^2Y both exceeding 0.99 (Figure 8.42). According to default cross-validation settings in the SIMCA software only five components are fully significant. We can see that most of the Y-variation has been modeled after three components. Typically, however, in NIR the predictive power resides in higher-order components.

By testing external predictivity against the 24 (2 times 12) prediction set observations, it was observed that ten components were optimal. Figures 8.43 – 8.45 show the results from the external predictions. As seen, there is excellent agreement between measured and predicted birch, pine, and spruce proportions in the sawdust blends.

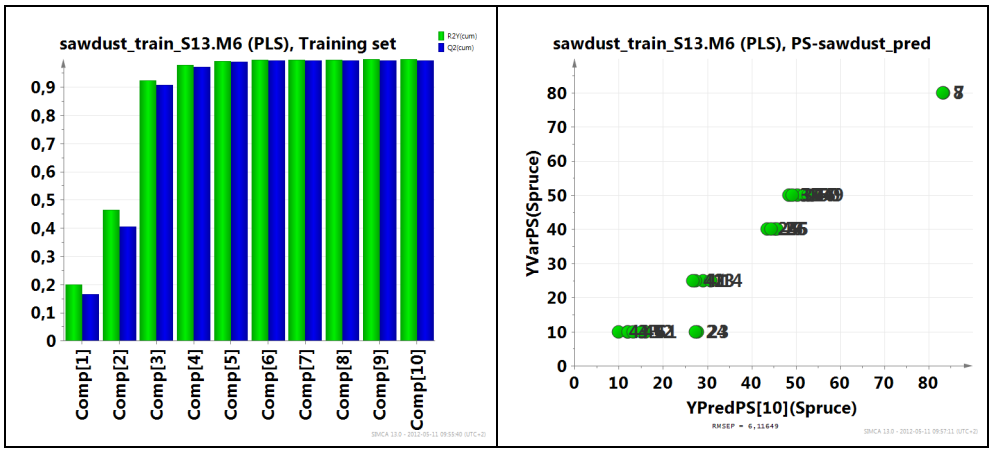


Figure 8.42: (left) Summary of fit plot of the PLS model of the SAWDUST training set. Ten components were found necessary using external validation. Figure 8.43: (right) Agreement between observed and predicted proportions of spruce in sawdust. Note that only prediction set samples are plotted.

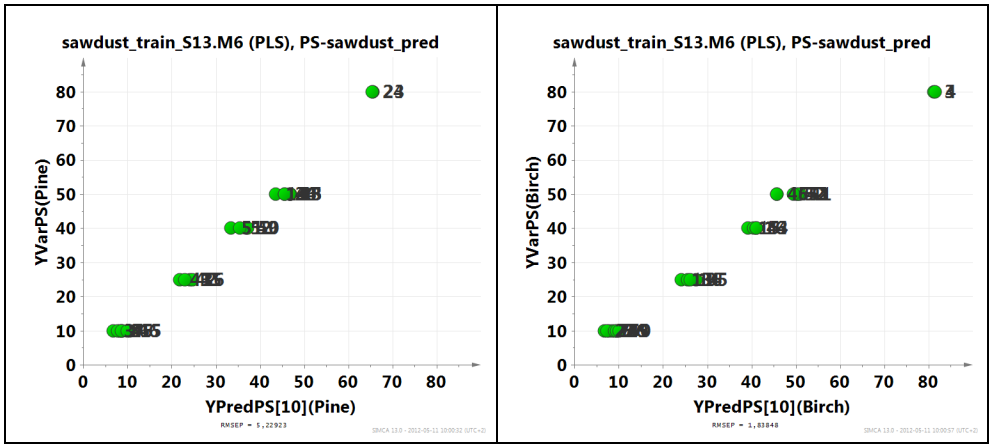


Figure 8.44: (left) Agreement between observed and predicted proportions of pine in sawdust. Figure 8.45: (right) Agreement between observed and predicted proportions of birch in sawdust.

8.6.4 OPLS modeling

The complementary OPLS model also gave five components according to cross-validation, two predictive and three orthogonal components, having the same overall statistics as the above PLS model. The fact that PLS and OPLS give the same number of components suggest the cross-validation rules in the SIMCA software to be well tuned. Moreover, obtaining two predictive components is according to expectation, since the rank of the Y-matrix (the mixture proportions) is two. A comparison of the PLS scores and the OPLS predictive scores plotted in Figures 8.46 and 8.47 clearly demonstrates the ability of OPLS to focus on the predictive variation in the spectral data; the triangular shape of the underlying mixture design is almost perfectly reproduced by the predictive scores of the

OPLS model (Figure 8.47). According to the OPLS model 31.2 % of the spectral variation is predictive to Y and 68.1% is orthogonal to Y.

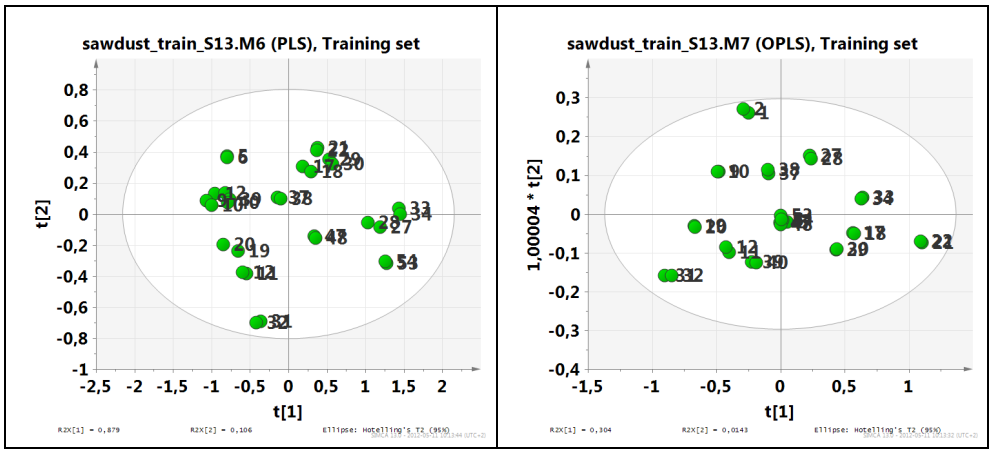


Figure 8.46: (left) PLS score plot of the first two components. Figure 8.47: (right) OPLS score plot of the two predictive components.

8.6.5 Summary of SAWDUST application

The conclusion of the SAWDUST application was that NIR is appropriate for modeling varying proportions of birch, pine, and spruce in sawdust. This demonstrates the potential utility of NIR for monitoring wood fiber composition in industrial applications.

The use of mixture design for training set composition (Figure 8.48) resulted in a representative and informative training set of sawdust blends. By selecting prediction set samples not completely identical to the training set samples (Figure 8.48), it was possible to validate the predictive ability of the calibration model in a realistic manner.

8.7 Questions for Chapter 8

1. What is calibration?
2. Discuss some of the limitations associated with traditional calibration!
3. Why is multivariate calibration better?
4. Give a short geometric description of multivariate calibration!
5. What are the main steps in multivariate calibration?

8.8 Summary and discussion

This chapter introduced multivariate calibration. The traditional univariate approach to calibration has been briefly reviewed and its limitations described. Traditional calibration suffers from selectivity, precision and accuracy problems and restricted model diagnosis capabilities. These limitations are overcome by the use of multivariate calibration because it uses a combination of signals from many wavelengths to model the data more efficiently.

It is, as always, important to have a good understanding of the problem at hand, i.e. which Xs to use, which Ys to use and which numerical ranges of Y (concentrations, levels, proportions) are appropriate. It should also be remembered that having many Y-variables is not a problem. In fact, multivariate Y-data provide more information and they can be separately analyzed by PCA. It is much more important to select a good, representative and diverse calibration set. This can be achieved using proper design and sampling in both X and Y [Flåten, *et al.*, 2004; Jonsson, *et al.*, 2004; Svensson, *et al.*, 1997]. The main principles are described in the following sections.

8.8.1 Two calibration situations

We recognize two main multivariate calibration situations:

(1) Reference samples can be prepared with desired levels of analytes and interferents.

(2) Samples are simply collected from the system and analyzed by a reference method.

Within these two types, other situations exist, depending on the number of analytes and interferents, and whether DOE and/or sampling strategies are applicable or not. The situations relating to point (1) above are discussed in Sections 8.8.2–8.8.5 and those relating to point (2) in Sections 8.8.6–8.8.8.

8.8.2 Reference samples with one or two constituents

A design in one Y-variable is appropriate when the reference samples contain:

(i) one analyte,

(ii) one analyte + interferent (summing to 100%),

(iii) two analytes (summing to 100%).

Here, it is recommended to use a multi-level design variable with 7–9 concentration levels [see, e.g., Berntsson, *et al.*, 2000; Berntsson, 2001]. One example is the multivariate calibration set discussed in Chapter 18 in connection with PAT (process analytical technology). In that application, two powders were mixed in binary blends, where the proportions were varied from 0% to 100% in steps of 10%.

8.8.3 Reference samples with multiple constituents

Section 8.8.2 describes a special case in that only one Y-variable is needed to specify fully the composition of the reference samples in the calibration set. However, when dealing with several constituents, several Y-variables are needed. The induced change in the concentration/level/proportion of each constituent defines a separate Y-variable. It is common for factorial and fractional factorial designs to be used in such cases [see e.g., Moberg, *et al.*, 2000; Moberg and Karlberg, 2001]. These types of design are applicable when there is a solvent or bulk constituent, uninteresting per se, which makes it possible to vary the levels of each analyte independently of all the other analytes. A mixture design is applicable when the constituents add up to 100 % (see Section 8.8.4).

8.8.4 Multiple constituents varied via mixture design

When the amounts of the constituents sum to 100% it is no longer possible to use an orthogonal design, such as a factorial or fractional factorial design, because the levels of the constituents cannot be varied freely. Instead a mixture design is appropriate [see e.g., Dyrby, *et al.*, 2004]. For instance, in the SAWDUST dataset (see Section 8.6), the proportions of birch, pine and spruce sawdust were varied using a simplex centroid design to create a diverse and representative calibration set (Figure 8.48). A complementary mixture design was set up for the prediction set (Figure 8.48). The latter design was deliberately constructed to be slightly different to the calibration set design to allow a more realistic assessment of predictive power to be made.

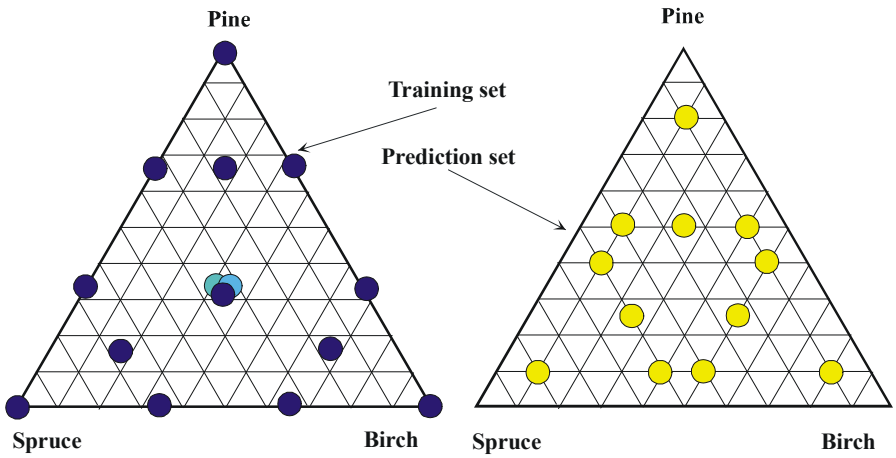


Figure 8.48: Spread of samples in the training set and prediction set in the SAWDUST application.

8.8.5 Combined design in constituents and process conditions

A way of making the calibration model robust to batch to batch variation in the main constituents (or other undesirable source of variation) is to make simultaneous changes to the levels of the constituents *and* the settings of the process factors. A good example of the power of this approach is provided by Gottfries and co-workers [Gottfries, *et al.*, 1996]. NIR transmission data, recorded in the range 800-1400 nm, were used to derive a calibration model capable of predicting the amount of API (metoprolol) in tablets with a precision of 1.05 mg. The design used was a mixed full factorial design in six factors and 400 runs (experiments); two factors were varied at five levels and the other four factors at two levels:

- Amount of API (metoprolol), 5 levels
- Tablet height, 2 levels
- Batch of metoprolol pellets, 2 batches
- Batch of microcrystalline cellulose, 2 batches
- Amount of water in tablet, 2 levels
- 5 tablets per batch

8.8.6 Samples are collected from the system and analyzed by a reference method

It is not always possible to use DOE to create calibration sets. In many cases, one must rely on a sound sampling strategy instead. Samples are collected from the system or process and analyzed by some reference method. Many samples, on the order of 50-100, or more, are needed to ensure that all relevant properties are adequately spanned [Wold and Josefson, 2000]. The exact selection procedure will of course depend on the specific context, e.g. whether it is to be driven on a batch to batch basis, or perhaps by geographical location. The former approach is typical in industrial applications while the latter often applies when dealing with environmental samples.

8.8.7 Location-based sampling

When dealing with “natural” samples, for example environmental or agricultural samples, the sampling strategy will inevitably be based on the geographical location of sampling sites (Figure 8.49). Furthermore, if the samples are of different varieties (not necessarily linked to geographical location), e.g. of wheat, all the different varieties must be included in the calibration set. Alternatively, different calibration models may be built for each variety.

If the calibration model is applicable to a finite set of locations, it is sufficient to select samples from those locations only, ensuring that each area is well covered and that samples are suitably divided into calibration and test sets. A simple way of doing this is to sort the

data according to a Y-variable and allocate alternate samples to the calibration and test sets. The multivariate design approach described in Chapter 7 may also be applicable, provided that multivariate characterization data are available for each sampling location.

However, if the calibration model has wider applicability and should also apply to new locations, then including samples from the new locations in the test set is a minimum requirement. Interesting developments, based on DOE approaches, to soil sampling planning in traceability studies for classification of Italian olive oil and red wine, are pursued by Marina Cocchi and her research group at Modena University in Italy.



Figure 8.49: Distribution of sampling sites in an air-pollution investigation. The levels of six persistent organic pollutants were determined, by sampling pine needles at eight growth locations in Europe. The sampling sites (marked by stars) are spread out according to a north-south gradient.

8.8.8 Batch-based sampling

Manufacturing processes can be sampled in a similar way to the environmental situation, but the sampling is based on batches rather than geographical locations. The calibration set should include samples from a representative set of batches and contain samples capturing a spread of time and a spread of any chemical, physical, biological or engineering properties that influence quality. The exact way of specifying the calibration set depends on the experimental context. Basically, it depends on whether one can sample to get a good spread of X and Y or not.

When the reference samples cannot be prepared with the desired levels of the constituents, sampling to get a good spread in Y is required. If only one Y-variable is involved, sorting according to the Y-variable, and allocating alternate samples will produce calibration and test sets with similar ranges of Y. With multiple Y-variables, multivariate design based on PCA and DOE can be used to select diverse and representative sets.

Sorting according to the Y-variable is recommended when data are acquired over time. This breaks down the inherent auto-correlation structure among adjacent samples and makes cross-validation more realistic. One should note however that, while this is a recommended approach for model building, it does not necessarily correspond to the optimal long-term strategy for process monitoring.

Multivariate design can also be applied to the X-data in order to get diverse and representative samples for (i) the calibration set and (ii) continued experimentation. Sometimes, however, it is not practical to use DOE to create the calibration set. In particular, consider complex industrial applications such as the use of NIR for determining the kappa number of pulp or the use of fluorescence spectroscopy to predict the ash content of sugar. Here, in order to understand how the process conditions vary, representative empirical data are gathered by monitoring the process over a long period of time. Such process data must then be split into representative calibration and test sets.

Finally, it must be understood that it may not always be possible to build good calibration models of a manufacturing process simply because the process does not vary sufficiently. This situation occurs because industrial processes are normally run within very tight specifications to ensure that useful product is manufactured. As a result, sample quality will follow a very narrow distribution. It is necessary to perturb such processes using DOE principles to generate sufficient variation. If this is not possible, laboratory or pilot plant scale analogues can be used to create sufficient quality variation.

An alternative approach is to use multivariate classification on the spectral data according to the SIMCA methodology. This approach is outlined in Chapter 10.

9 Multivariate process modeling

9.1 Objective

Multivariate characterization and multivariate calibration are two important fields of application of multivariate projection methods. These fields were discussed in the two foregoing chapters. In Chapter 9, we shall outline how PCA and PLS may be used in a third major application area, multivariate process modeling. Three examples, called CUPRUM, PROC1A and SOVRING, will be used to highlight the main steps of multivariate process monitoring.

9.2 Introduction

Chemical processes are often monitored at frequent time intervals, giving rise to data bases with many variables and many observations. On numerous occasions, PCA and PLS have been shown to work efficiently with process data [see, e.g., MacGregor and Kourti, 1995; Tano, 1996; Tano, 2005; Zeaiter, *et al.*, 2006; Liebmann, *et al.*, 2010; Butz, *et al.*, 2011; Viteri, *et al*., 2012; Bu, *et al.*, 2013]. This is because they can deal with the intrinsic problems of process data, that is, problems related to (i) dimensionality, (ii) collinearity, (iii) noise and (iv) missing data.

The multivariate approach is appropriate for several reasons. Firstly, process drifts are not always detected by looking at the variables one at a time. Rather, a drift in a process is often expressed by a group of correlated variables, whose values move up and down together. Thus, the information resides in the variable correlation pattern, rather than in the individual variables. Secondly, as soon as samples show up which break the general correlation structure, these can easily be spotted by the various diagnostic graphs (explained later). In addition, the use of several correlated variables will result in a strengthened latent "signal" with improved signal-to-noise ratio.

Finally, the multivariate approach based on PCA and PLS offers a number of informative diagnostic tools, which facilitate the identification of *assignable causes*, i.e., they help us to understand how and why a particular process event occurs. With such information available, it becomes possible to take corrective action in order to prevent future occurrences of the out-of-control conditions.

In this chapter we will give an introduction to multivariate process modeling. First we describe the objectives of multivariate process modeling, then the nature of process data, and finally we address three different applications. It should be noted that more elaborate applications of process modeling are described in Chapters 15, 16, and 17. Chapter 15 discusses how control chart theory can be implemented in a multivariate manner in statistical process analysis. This is called multivariate statistical process control, MSPC. Similarly, Chapter 16 addresses control charting and modeling of batch process data, an approach which we call batch statistical process control, BSPC. Finally, Chapter 17 deals with multivariate time series modeling, where the aim is to model and understand the auto-correlative behavior among observations.

9.3 Multivariate process modeling

9.3.1 Objectives

Multivariate process modeling is a multi-faceted discipline. It can be used for supervision of existing processes, improvement of existing processes, and development of new processes. Supervision of existing processes is accomplished by monitoring process data and referencing these against a model based on historical data. In order to improve the performance of existing processes, models developed on historical data are interrogated to uncover the most influential process parameters. In the next step, DOE in these important factors may facilitate the detection of better process conditions. Finally, multivariate process modeling coupled with DOE might be used when developing maintenance strategies for new processes.

Regardless of whether new or existing processes are to be modeled, the objectives of the data analyses are the same, namely (i) *monitoring the state of the process*, (ii) *understanding the relationship between factors (X) and responses (Y)*, and (iii) *optimization to achieve optimal process behavior*. Monitoring the state of a process multivariately (process status visualization, PSV) gives the process operator advanced warning of when the process shifts from one condition to another, and suggests the probable cause of the diagnosed process change.

Further, if there are both input variables (factors, X) and output variables (responses, Y), these can be analyzed together by means of PLS. With a PLS model it is possible to understand which X-variables are influential for a particular Y-variable, which X-variables provide similar and unique information about the current state of the process, and how the Y-variables are associated with each other. An understanding of the dominant relationships among the input and the output variables makes it possible to carry out process optimization by introducing appropriate modifications in the most important factors.

9.3.2 Process data

Fifty years ago, a process operator had access to only a limited number of process variables, like temperature changes, pressure fluctuations, and material flows. This enabled data to be overviewed simply by looking at a time series plot of each individual variable. A time series plot displays how the measured numerical values change as a function of time. Today, however, process measurements are made regularly and much more extensively than in the past. Figure 9.1 shows typical processes. The masses of process data which are measured by computers and electronic sensors are usually stored in extensive data-bases. Some properties and challenges of such highly multivariate data were discussed in Section 2.3.



Figure 9.1: Typical processes where signals are measured within the different unit operations and between them.

Processes can be divided in two main categories: *continuous* processes and *batch* processes. In a continuous process, the machinery is operated continually and is not stopped until external factors require it (service & maintenance, holiday, …). Continuous processes can be operated under *steady-state* or *dynamic* conditions. When steady-state applies, observations (time points) should theoretically show a low degree of interdependency. Conversely, when a process shows dynamic behavior, with some memory and drift, observations are usually highly correlated over time.

In contrast to a continuous process, a batch process lasts for a finite time. The raw materials are processed until a certain pre-determined level of quality, or grade of product, has been obtained. For batch processes, a distinction between steady-state and dynamic behavior can also be made. In the steady-state perspective, only variation *between* different batches is considered. But when dynamic properties are to be explored, the variation over time *within* batches is also analyzed.

In addition to dividing processes into two categories, we may also divide the variables into different blocks. Variables might be: (i) manipulated, controlled process variables (X), like temperatures, flows and pH; (ii) passively observed process variables (X), like raw material changes, humidity, temperature, etc.; (iii) intermediate process variables (X), like viscosity or concentration at exit of reactor 1, which becomes input to reactor 2; (iv) quality variables (Y). Most X-variables are measured continuously, whereas samples are usually withdrawn to measure the quality (Y-) variables at discrete times.

The main purposes of measuring the data on the process, its input and environment are to:

(i) Provide information allowing a better understanding of the process, relationships between different parts of the process, which chemical or other reactions occur inside reactors, and so on.

(ii) Analyze yield information about the state of the process, recognizing trends, groupings, peculiarities, etc., all to keep the process under proper control.

(iii) Discover how the output is affected by the process and input variables in order to improve product quality and minimize manufacturing costs, energy consumption, pollution, etc.

(iv) Analyze data to understand when preventive maintenance could have helped to avoid more costly unscheduled downtime.

9.3.3 Applications

Process datasets are not very different than data acquired from spectroscopy or drug design. Usually, there are a few latent factors that are expressed by the measurable variables, and observations may, or may not, be grouped. One characteristic feature of process data, however, is the strong auto-correlative behavior of observations that is usually present.

PCA and PLS produce scores which can be used to overview a process, trace its evolution, and detect groupings and outliers among observations. Strong outliers are found by considering the scores, and moderate outliers by inspecting the model residuals (DModX/DModY). In addition, the resulting loadings, weights, VIP, etc., may be used for model interpretation, in order to understand how to set X to get a desired Y.

In order to demonstrate the wide variety of process applications, we shall study three examples. The first application (CUPRUM) exemplifies an overview of process quality data, the second example (PROC1A) shows how process monitoring might be carried out, and the third example (SOVRING) highlights how relationships between process variables (X) and quality variables (Y) can be modeled with PLS.

9.4 Example: CUPRUM

Our first example of multivariate process modeling is called CUPRUM. An electrolysis process at Boliden AB (Skelleftehamn, Sweden) was sampled during one year of copper production. It is to be regarded as a continuous process. Two copper samples were taken each day, creating a data base with 730 samples (observations, rows). Each copper sample was investigated for impurities, and the levels of eight metal species (in ppm), viz., Ag, Ni, Pb, Bi, Sb, As, Te, and Se, were recorded (variables, columns). For each species Boliden defined an upper control limit. In the manufacturing process, each cathode was left in the electrolyte for 10 days. During this period, the amount of copper was allowed to continually increase until finally harvested.

9.4.1 Quality control with a univariate index

One way to monitor the copper quality is to use a univariate quality index. Such an aggregate index may be calculated as a weighted sum of the eight measured variables. The tolerance limits set by the company were used as weights, and the resulting index is called the total analysis index (TAI). If TAI is below 8, the copper sample is classified as of acceptable quality; TAI greater than 8 indicates a rejected sample. Figure 9.2 shows a time series plot of TAI, with 19 samples having TAI > 8.

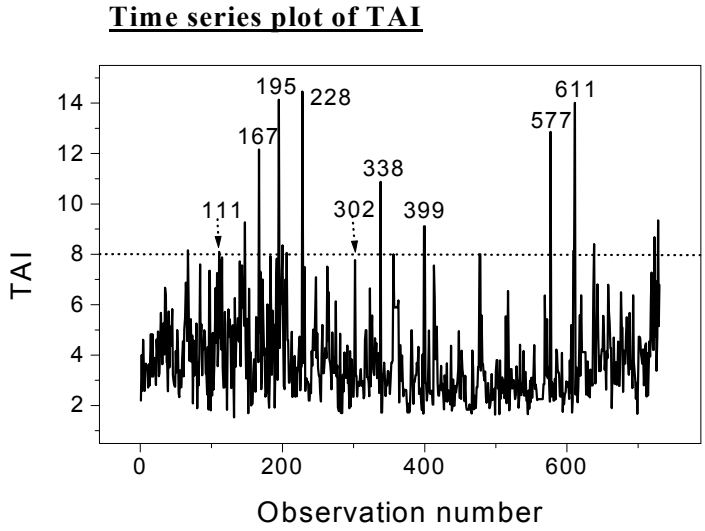


Figure 9.2: A time series plot of the total analysis index (TAI). Nine samples are marked for reference purposes (see text).

The two time points (samples) 111 and 302 both have similar TAI values. TAI for 111 and 302 are equal to 8.1 and 7.8, respectively. According to this scale both samples are of comparable (low) quality (TAI ranges from 1.5 to 14.5). However, Figure 9.2 does not reveal why these samples are similar in quality, nor what the relationships among all the process points are.

9.4.2 A better approach - PCA

With PCA, the information content in all eight quality variables is considered at the same time. The PCA modeling resulted in a model with two components explaining 67% (R^2X = 0.67) and predicting 40% (Q^2X = 0.40) of the total sum of squares. This means that the score plot displayed in Figure 9.3 is a good summary of the copper sample data, as it represents more than two-thirds of the total variation. Two components also means that only one quality index, like TAI, is insufficient. The two samples discussed (111 and 302) are marked. Interestingly, and in contrast to the TAI-scale, these two samples are clearly modeled as having markedly different quality properties.

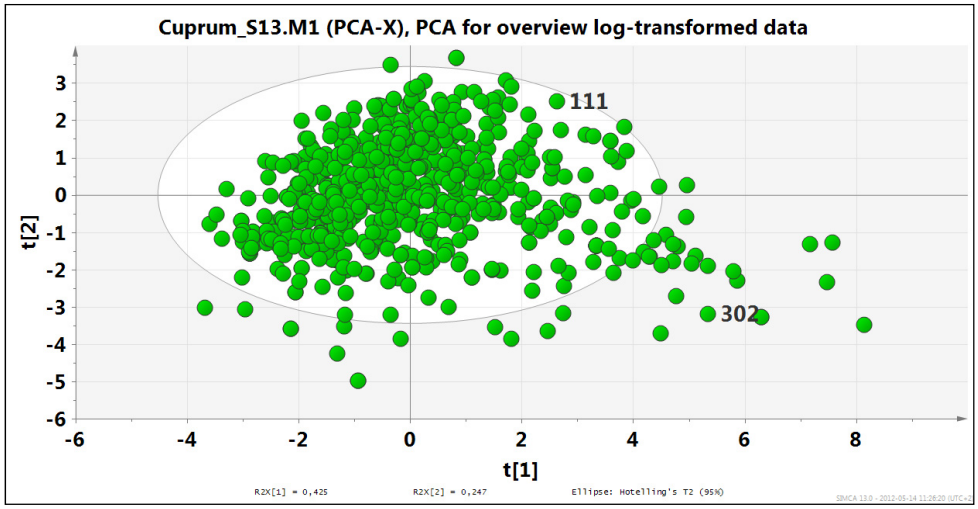


Figure 9.3: PCA score plot with t_2 versus t_1 accounting for 67% of the total variation. Each copper sample is shown as a unique plot mark. Samples 111 and 302 are highlighted for reference purposes.

To understand and interpret the distribution of the copper samples in Figure 9.3, we can examine the corresponding loading plot shown in Figure 9.4. The eight quality variables are strongly grouped in two clusters, indicating two distinct impurity types. The first principal component reflects "average" impurity and the second principal component reflects deviations from this "average" level. Thus, sample 111 has comparatively high levels of the trace impurities Sb, As and Bi (all group 5a elements in the periodic table !), and low levels of the species Ni, Pb, Se, Ag and Te. Sample 302 exhibits the inverse impurity pattern. Such different impurity patterns correspond to copper samples of differing quality. This quality difference – revealed with PCA – was undetected by the TAI-scale, because the combination of eight variables into one single scale (TAI) resulted in a severe loss of information. The TAI-scale does not consider the correlation structure among the eight quality variables.

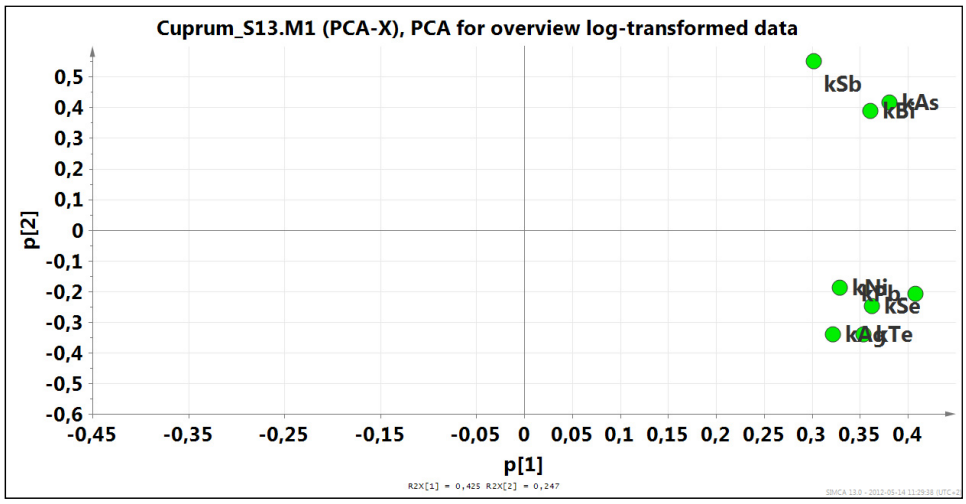


Figure 9.4: PCA loading plot corresponding to Figure 9.3. Each single quality variable gives rise to one plotted point. The interpretation suggests that the first principal component reflects "average" impurity, because all loadings are positive. The second principal component models deviations from this "average" impurity level.

9.4.3 Zooming-in on highlighted samples

The loading plot gives an overall picture of the variable correlation structure. In order to "zoom in" and investigate the profiles of the variables of samples 111 and 302, we may use another graphical tool, the *contribution plot*. This diagnostic tool may be thought of as positioning a magnifying glass over samples 111 and 302. The contribution plot is of

general applicability in multivariate data analysis, but has found particular acceptance in process data analysis. It makes it possible to compare two samples with each other, or to compare an interesting sample with a reference sample, typically the average process point.

Figures 9.5 and 9.6 show the contribution plots of observations 111 and 302. These show that sample 111 has elevated levels of Bi, Sb and As compared with the average sample, and that sample 302 has higher values of Ag, Ni, Pb, Te and Se (and to some extent As) than the average sample. From the process operator's perspective, this type of contribution plot greatly facilitates process fault assessment.

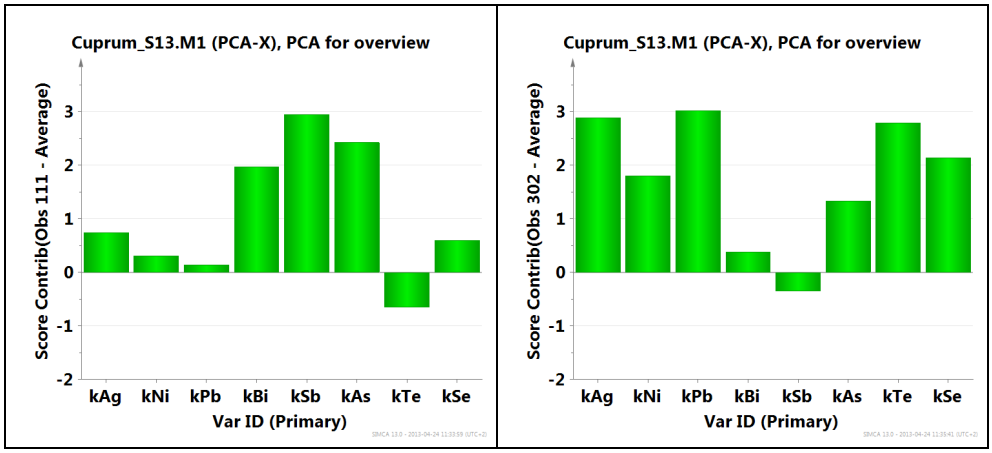


Figure 9.5: (left) Contribution plot (scores-mode) showing the deviation between observation 111 and the average process point. Sample 111 has elevated levels of Bi, Sb, and As. Figure 9.6: (right) Contribution plot (scores-mode) showing the deviation between observation 302 and the average process point. Sample 302 has higher values of Ag, Ni, Pb, Te and Se (and to some extent As) than the average sample.

We realize from Figures 9.5 and 9.6 that contribution plotting is useful in process analysis for understanding the difference between two observations in a score plot. Moreover, a group-based extension is used when comparing the difference between the averages of two groups (or classes) spotted in a score plot. Interestingly, contribution plots can also be formulated for differences in the residuals (DModX). We will postpone further discussion of the CUPRUM application until later, and use the next section to give an introductory description of the basis of contribution plotting.

9.4.4 Contribution-based diagnostics

In the context of process analysis, plots of scores and residuals (DModX) often show sudden "jumps". A practical way to understand such shifts is to use contribution plots [MacGregor and Kourti, 1995], which reveal the set of process or quality variables which make the greatest contributions to the observed "jump". Contribution plots can be calculated in two ways, relating to "jumps" in score plots or DModX plots.

Differences in scores: In this case the contribution plot is used to identify the changes in the variables between two observations, or between one selected observation and the average process observation, or between the averages of two groups. This is computed as

$$\textit{Contribution (scores)} = \Delta X * \textit{weight} \qquad \text{(eqn. 9.1)}$$

where the default weight is the component loading, p.

Deviations in DModX: Here the contribution relates to residuals which, in turn, point to the variables. The contributions are calculated as

$$\textit{Contribution (DModX)} = e_k * \textit{weight} \qquad \text{(eqn. 9.2).}$$

Here, e_k, represents variable residuals and the default weight is the square root of the explained sum of squares for each variable.

In summary, contribution plots will indicate which variables are the greatest contributors to a certain process shift. Although they will not explicitly indicate the cause of the shift, they

provide a good insight into plausible causes or variables related to the causes. This is a good starting point for further fault analysis.

With this short explanation in mind, we now return to the CUPRUM example.

9.4.5 Plots of T^2 and DModX are more informative than TAI

In order to understand whether there are outliers in the data, we can use the T^2- and DModX-statistics (see Section 3.5). These parameters can be plotted in a joint control chart as functions over time (Figure 9.7). Such a combined control chart makes it possible to survey how the process points develop as a function of time, with particular emphasis placed on detecting strong and moderate outliers.

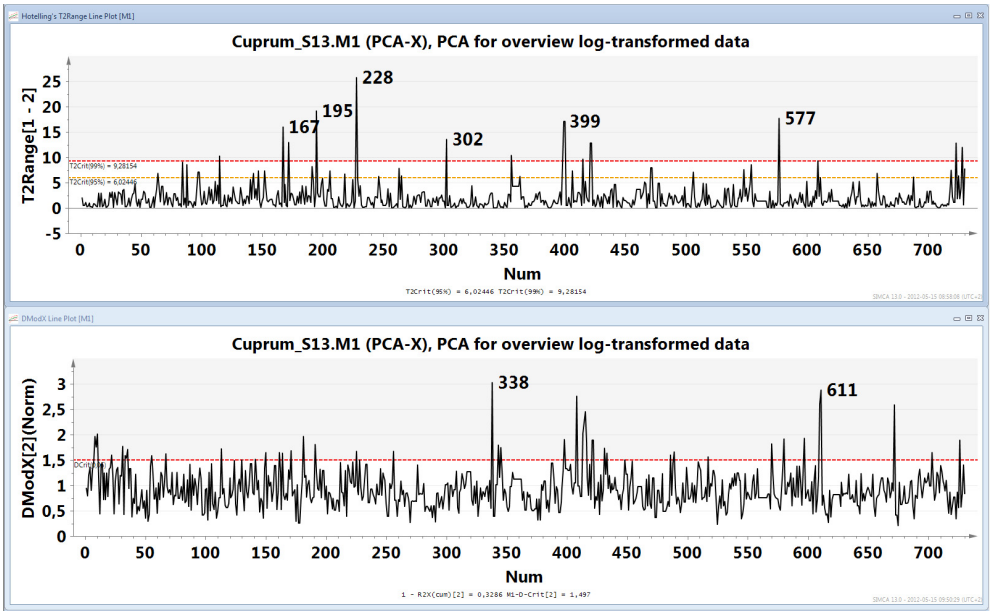


Figure 9.7: Combined control chart with T^2 and DModX versus time order. Tolerance levels corresponding to 95% and 99% are given for T^2, and the critical distance for DModX is plotted. Seven samples are marked for comparison with Figure 9.2.

The top part of Figure 9.7 shows how T^2 evolves over time, and the bottom part how DModX changes over time. Here, T^2 is a summary of all eight quality variables and all model dimensions. It will detect strong deviations in the systematic part of the data (strong outliers). Warning and action limits can be set at, for instance, the 95% and 99% levels, respectively. We can see that the copper quality has deviated substantially from the nominal area in at least six (maybe even ten) instances. The supplementary DModX control chart displays the model residuals, and can be used to identify deviations from the overall correlation structure (moderate outliers). In summary, by using the T^2 and DModX plots at the same time, we can see if the copper quality tends to go out-of-control occasionally, or if a more persistent trend is establishing itself. In the current application, no such conspicuous longer sequences of observations can be seen.

It is interesting to compare Figures 9.2 and 9.7, where the same seven samples (# 167, 195, 228, 338, 399, 577 and 611) are marked. These samples are, according to the TAI-scale, among the most impure copper samples (Figure 9.2). The combined control chart in Figure 9.7 also identifies these, but differentiates them into two categories. Five samples are abnormal according to T^2 and two according to DModX. Thus, the former five samples (167, 195, 228, 399 and 577) conform to the dominating correlation structure, whereas the latter (338 and 611) break the general correlation structure. This type of information is not available in overall aggregates, like TAI. Further interpretation of the differences between these two categories may be obtained in, for instance, contribution plots. For reasons of brevity, however, we do not display such plots.

9.4.6 Summary of CUPRUM application

In the copper example, the use of the TAI index was not very successful. The two samples 111 and 302 were claimed to have similar quality, there being no indication of their completely different impurity patterns. In contrast, the PCA suggested that these samples were not of comparable quality, because they were situated far apart in the score plot (Figure 9.3). The loading plot (Figure 9.4) provided clues for the interpretation by indicating two kinds of impurity profile. This is possible because PCA models the correlation structure among all quality variables. The eight quality variables migrate up and down in clusters and do not vary independently of each other. This information was not shown by the TAI-scale.

9.5 A second example: PROC1A

In the previous example (CUPRUM) the process never really went out-of-control; only occasionally were some samples of low quality. However, our second example (PROC1A) illustrates what may happen when process conditions become severely out-of-control. This example deals with a chemical production plant manufacturing a polymer. It is a continuous process.

The dataset contains 33 variables and 92 hourly observations. The measured variables are distributed as seven controlled process variables (x1in – x7in), 18 intermediate process variables (x8md – xpen), and eight output variables (y1 – y8). All data are coded to avoid disclosing sensitive information.

A raw data plot of all 33 variables is given in Figure 9.8. It is known that the process went out-of-control after 80 hours, and eventually had to be shut down. This was not obvious from a raw data inspection (Figure 9.8).

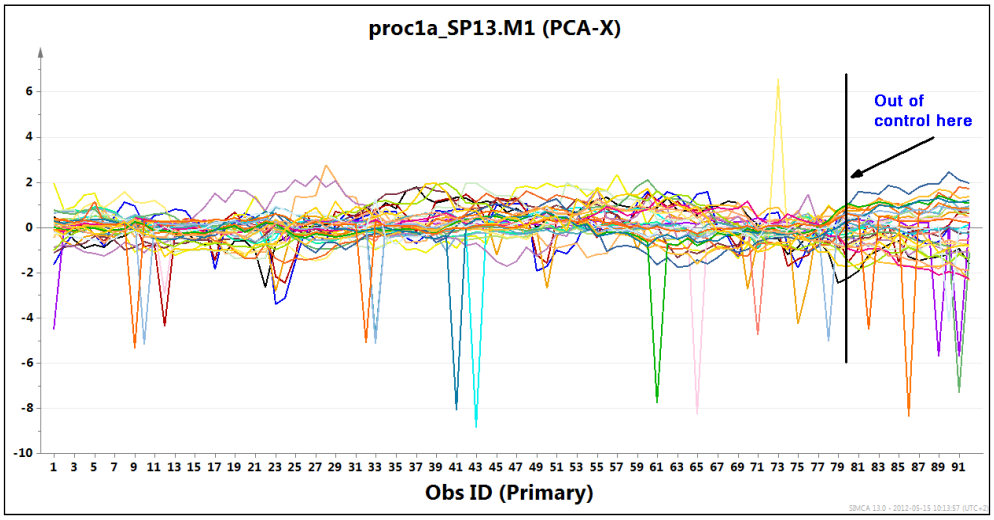


Figure 9.8: Plot of raw data of the PROC1A dataset. There are 33 trend curves superimposed on top of one another. The data have been scaled to better fit onto one plot. Simply plotting the raw data does not reveal that anything strange is happening after time point 80.

9.5.1 Establishing an overview of the process data

It is always recommended to start with an overview of the data. We used PCA and computed two components ($R^2X = 0.39/Q^2X = 0.27$), the scores of which are plotted in Figure 9.9. This score plot reveals a clear trend in the data. Close inspection of this plot shows that from observation 70 the process moves from the bottom of the graph steadily towards the upper left-hand corner. This movement indicates some kind of process upset.

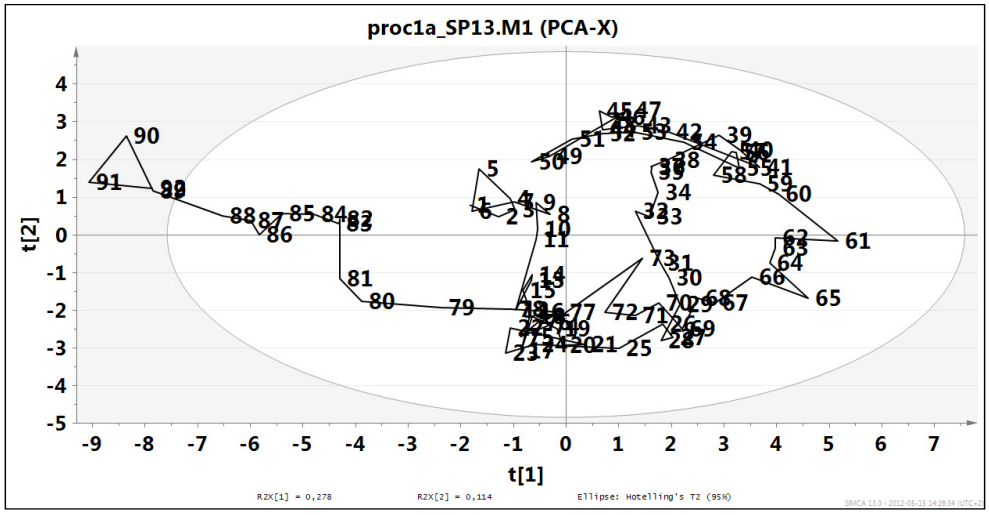


Figure 9.9: PCA t_1/t_2 score plot of the initial model of the PROC1A dataset. At the beginning, when the process behaves well, the observations wander around in a small region of the score plot. At the end, when the process behaves badly, a strong trend is expressed in the data.

9.5.2 Modeling the normal process behavior

According to the responsible process operator, the process was out-of-control from time point 80. However, the PCA above indicated process problems already from time point 70. Hence, we conclude that normal behavior of the process is seen in observations 1-69. Accordingly, a PCA model was fitted to the first 69 samples. Three components were significant according to cross-validation and the explained variation was 51% ($R^2X = 0.51$).

We can now use the historical data to define limits within which the process is in control. One type of control limit is provided by Hotelling's T^2, which models the deviation of the process in the PCA model hyperplane (Figure 9.10). As shown by Figure 9.10, the first 69 observations are inside the 95% tolerance limit. A second type of control limit is indicated by DModX (Figure 9.11), which monitors the deviation of the process perpendicular to the PCA model hyperplane. As shown by Figure 9.11, there are quite a few samples which exceed the critical limit in DModX. Usually, deviating observations in DModX indicate minor process upsets not captured by the model. However, if many observations in a sequence show consistently high DModX-values, then that is an indication of a "new", persistent process event. Thus, in the time interval 1-69, the process is fairly stable, and no new process event is developing.

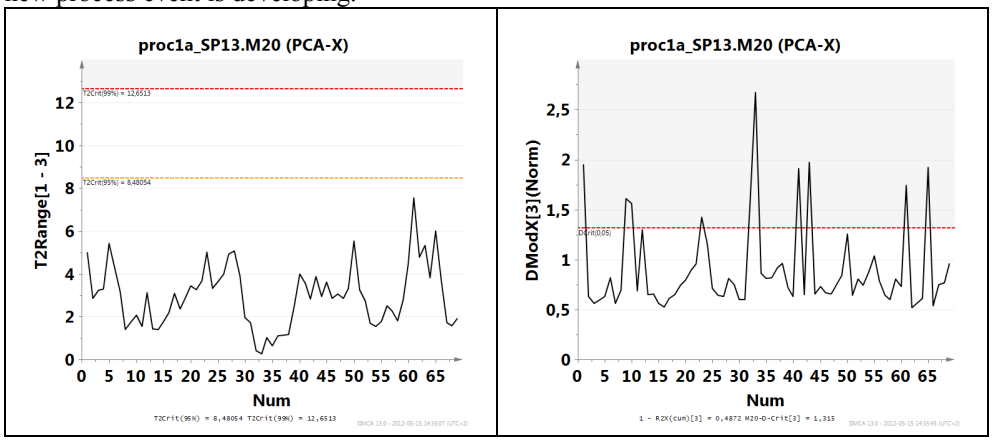


Figure 9.10: (left) Hotelling's T^2 can be used to define a control limit in the PCA model hyperplane. Obviously, among the first 69 observations, all samples conform to the model and are found inside the 95% critical limit.
Figure 9.11: (right) DModX can be used to define a control limit in a direction perpendicular to the PCA model hyperplane. Approximately 10 observations exceed the critical distance, and might be classified as moderate outliers. However, the important feature to recognize is that there is no sequence of observations showing persistently high DModX-values. This indicates that in the time interval 1-69, the process is fairly stable, and no new process event is establishing.

9.5.3 Using the model to monitor the process

When a model has been established on data which represent normal process behavior, it can be applied to new process data. The new process measurements are fed into the existing model, projected onto the lower-dimensioned score space T, and their fit is evaluated in the light of control limits set by Hotelling's T^2 and DModX. By this procedure, it will be possible to detect deviating observations, time trends, clusters, and so on. Once a deviating observation, or group of observations, has been discovered, loading and contribution plots can be deployed to determine which variables contribute to any particular process shift. Hopefully, once it is known where to search for a solution, it is then an easy task to correct the problem in the process.

The model obtained was applied to the remaining 23 samples (observations 70-92), the results of which are summarized in Figures 9.12 and 9.13.

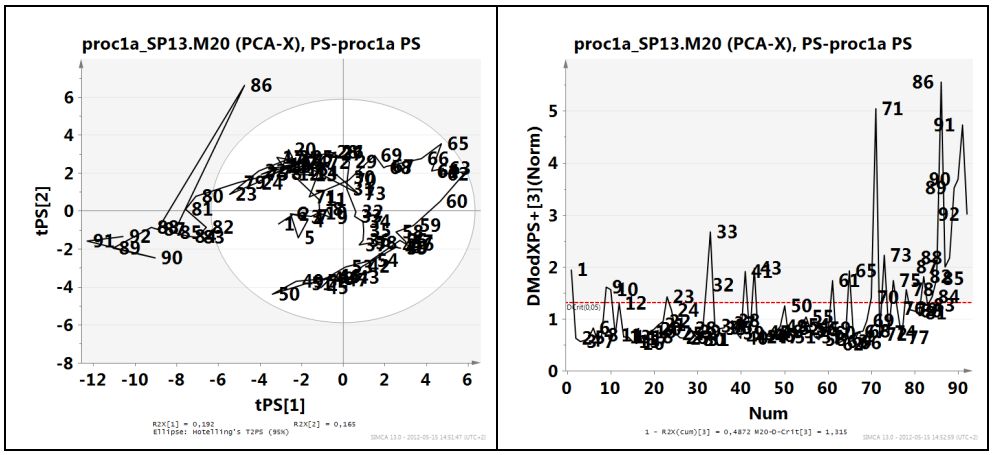


Figure 9.12: (left) PCA score plot of predicted score t_2 versus predicted score t_1. Projecting the new data onto the PCA model plane clearly indicates the process upset starting at time point 70 and becoming severe at time point 80. This is understood because samples 80-92 are positioned outside the T^2 control limit. Figure 9.13: (right) Distance to the model of the 92 observations. Most of the observations 1-69 used to train the model are inside or nearly inside the critical distance. However, observation number 33 is somewhat high and ought to be scrutinized more closely. The last observations do not fit the model well, and indicate that a new process behavior has become established.

Figure 9.12 displays a PCA score plot comprising the predicted scores of the first two dimensions. The first 69 observations were used to build the process model. Therefore, it is consistent with expectation that they are inside the T^2-control limit. However, as soon as the model is applied to the remaining observations (70 – 92), we can see that the last 13 (80 – 92) fall outside the control limit. This indicates that observations 80 – 92 are fundamentally different from samples 1 – 69, i.e., a new process behavior is emerging. It would be interesting to know in what way sample 80 is different from the average process point. This will be assessed in the next section.

Similar information is obtained from the plot of DModX (Figure 9.13). In general, most of the samples before time point 70 are below the critical limit, whereas the opposite is true for the samples after time point 70. Thus, according to DModX, the last thirteen test samples are different from the 69 "normal" samples. However, one sample that might usefully bear closer investigation is no. 33, as it displays the highest DModX among the training samples. This is investigated in the next section.

9.5.4 Identifying contribution from variables

In order to understand which variables contribute most strongly to an observed process change, we may use contribution plotting. One type of process change can be detected in the scores. As an example, we may take observation no. 80 (and all the ones thereafter), which deviates from normal behavior. A contribution plot related to observation 80 is given in Figure 9.14. This plot highlights which variables contribute to the difference between

observation 80 and the average process point. Evidently, in relation to the average process point, observation 80 deviates chiefly in variables x1in, xemd, xgnx, xoen, and xpen, by having substantially lower values in these than the average point.

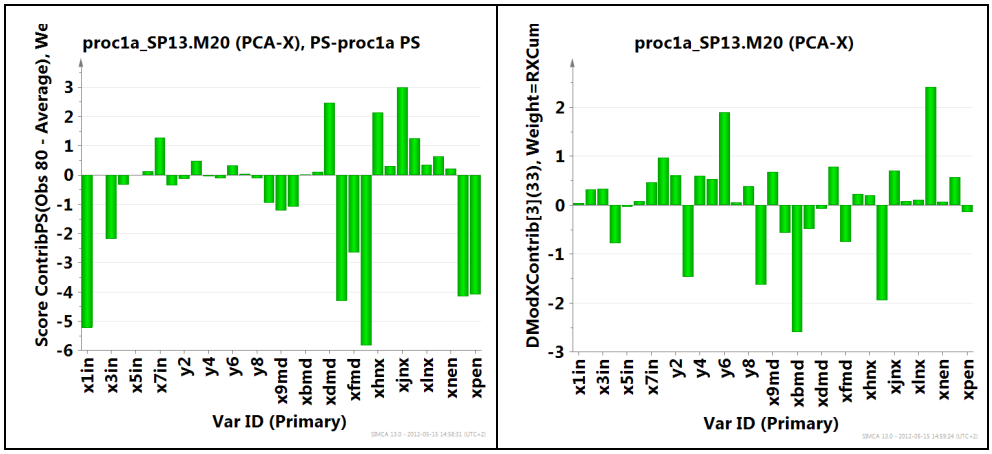


Figure 9.14: (left) Contribution plot in scores mode, displaying how observation 80 relates to the average process point. Primarily, observation 80 has much lower values in the five variables x1in, xemd, xgnx, xoen, and xpen. Figure 9.15: (right) Contribution plot in residual mode, showing why observation 33 exhibits a comparatively large DModX. Two variables, xbmd and xmen, contribute substantially to the large residual of sample 33.

The second type of process change can be evaluated in a plot of DModX. As an example, we may consider observation 33, which exhibits the largest DModX among the observations utilized to build the PCA model. The contribution plot shown in Figure 9.15 reveals which variables contribute to the high residual of observation 33. The two variables xbmd and xmen are the main contributors to this large residual.

9.5.5 Exploring time series properties of scores

In process applications the time series properties of the scores are also important. Because the scores represent weighted average trajectories of the original variables, exploring their time dependence will give good insight into trends and shifts in the process. Figures 9.16 and 9.17 represent time series plots of the first two predicted scores (t_1 and t_2). The process fluctuates up and down in these scores, but towards the end of the measuring period extreme values are observed for both scores.

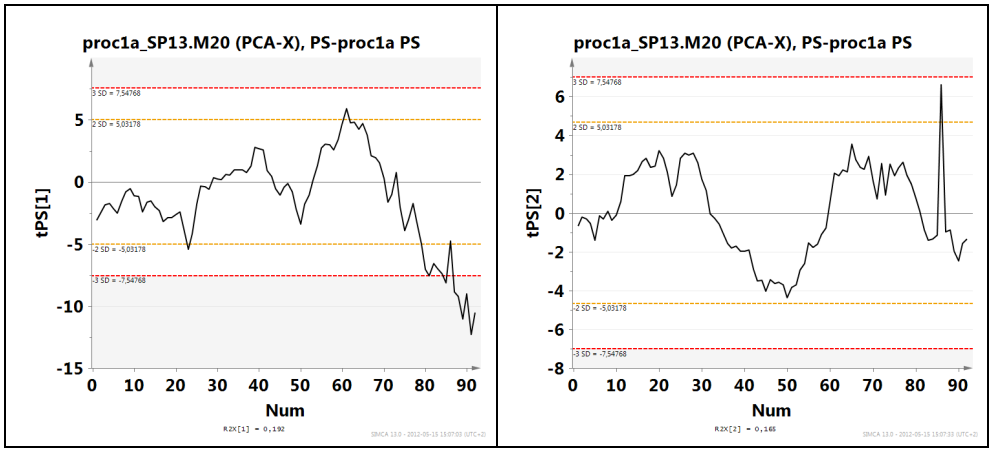


Figure 9.16: (left) Time series plot of predicted t_1. Figure 9.17: (right) Time series plot of predicted t_2.

9.5.6 Summary of PROC1A

The PROC1A application exemplifies how multivariate monitoring of a chemical production plant may be accomplished. The first step in the data analysis corresponded to an overview of historical process data, by means of principal component analysis. This indicated a clear process upset after observation 80. Then, we tried to decide on what would be regarded as normal process operating conditions. Based on process knowledge and the initial PCA overview, the first 69 observations were recognized as representing normal process behavior.

Subsequently, the monitoring PCA model was calculated based on the selected set of 69 observations. Cross-validation was used to determine that three principal components constituted the optimal model complexity. The control limits in T^2 and DModX were calculated, and used to evaluate the model conformity of test data (samples 70 - 92). In the monitoring stage, it was found that observation 80, and subsequent ones, were all projected outside the T^2 tolerance ellipse. A contribution plot showed that the problem was to be found in a number of variables, such as, x1in, xemd, xgnx, xoen, and xpen, whose values were all too low. Having acquired this information, it becomes easier for the process operator to apply corrective measures.

9.6 A third example: SOVRING

In the third example, we shall study how PLS can be used in process data analysis. PLS models relationships between blocks of variables, for instance process factors (X) and quality measurements (Y). Thus, with a PLS model it might be possible to understand how to modify the regulation of a process in order to improve product quality, minimize costs, and decrease pollution. As in PCA, the score vectors from PLS are new variables that summarize all process variables. They may be monitored separately. In addition, DModX and DModY summarize the unmodeled variation in the process, and they can be used to uncover new process events.

The example that we shall study is called SOVRING, and was kindly contributed by Kent Tano, a research engineer at LKAB, Malmberget, Sweden [Tano, 1996]. The data originate from a mineral sorting plant. In this process, raw iron ore (Ton_In) is fractionated into finer material by passing through several grinders. After grinding, the material is sorted and concentrated in several steps by magnetic separators. The separation flow is divided in several parallel lines, and there are also feedback systems to make the concentration of Fe as high as possible. The concentrated material is divided into two product streams, one called PAR and another called FAR, the latter of which is sold directly. The primary goal of this study was to establish which process factors were most influential, and to set up a prediction model that could be used in an on-line environment.

The dataset contains 572 observations (time points) and 18 variables. Twelve process factors were identified by Kent Tano, three controllable and nine of intermediate character (= 12 X-variables). The X-data were drawn from the process log on a minute-by-minute basis. The three controllable factors were explored using a composite design in 17 experiments. For each experiment, six response variables (= 6 Y-variables) were recorded. Each factor combination in the design was then preserved for 20 minutes to better capture normal process variation.

9.6.1 An overview PCA model

PCA was used on the entire dataset. The first two PCs were calculated ($R^2X = 0.79$, $Q^2X = 0.70$) in order to get an overview of the process. The scores of these components are plotted in Figure 9.18 and the concomitant loadings in Figure 9.19.

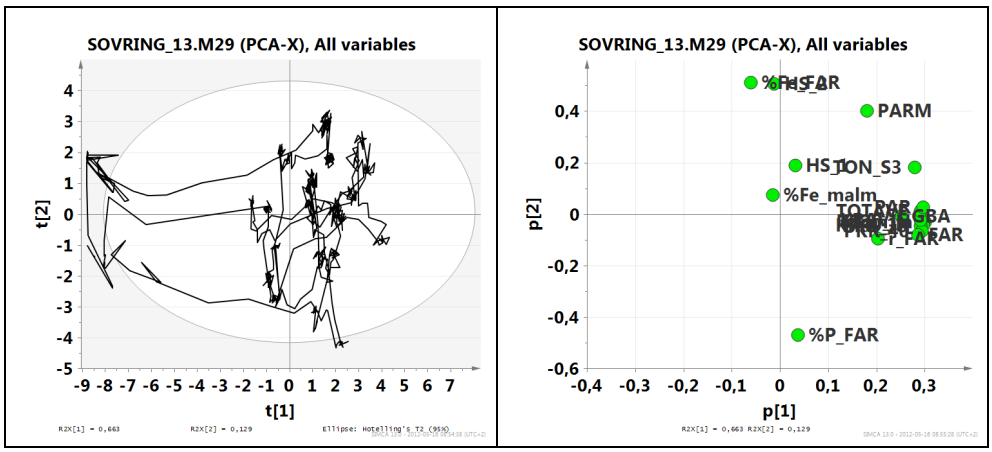


Figure 9.18: (left) PCA score plot (overview model, SOVRING) with t_2 plotted against t_1. Two minor clusters of process points deviate in the left-hand part. The two clusters to the left are due to no material flow. Figure 9.19: (right) PCA loading plot corresponding to Figure 9.18.

Apparently, there are two minor clusters of observations discernible in the left-hand part of the score plot. These deviating observations represent extreme and unwanted process conditions (no starting material). As a consequence, they should not be allowed to influence any model that is fitted to reflect normal process conditions.

9.6.2 A refined PCA model

As described in the foregoing section, chemically irrelevant samples were excluded from consideration. Furthermore, a selection of representative observations was made such that only 5 observations were kept for each of the 17 design runs, i.e., 85 observations were selected for refined modeling.

A new PCA was undertaken and four components were calculated ($R^2X = 0.90$, $Q^2X = 0.76$). Figure 9.20 displays a plot of the first two score vectors (t_1/t_2). Analogously, Figure 9.21 shows a plot of score vectors three and four (t_3/t_4). The former score plot shows a beautiful clustering of the 85 observations arising from the underlying composite design. The three center-points are well reproduced. The clustering is not as pronounced in the latter score plot, but the regularity of five points in each little cluster is still unmistakable.

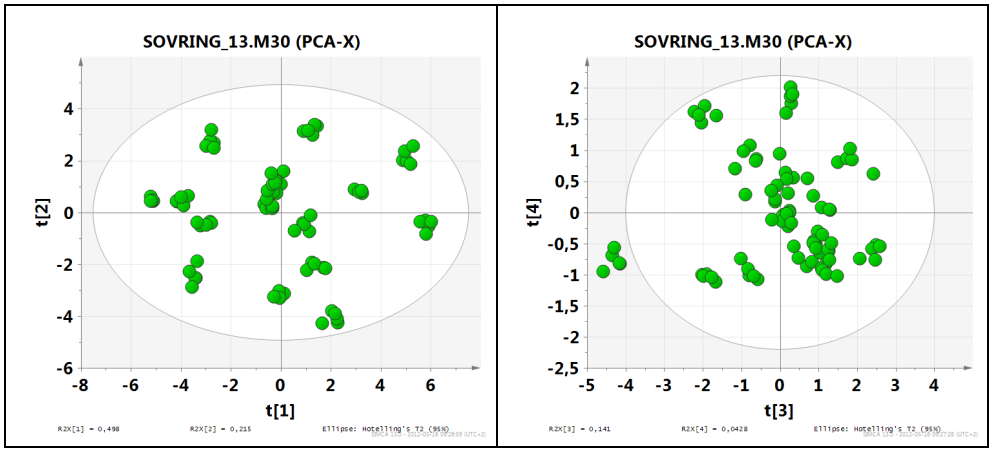


Figure 9.20: (left) PCA t_1/t_2 score plot (refined model, SOVRING). The sharp clustering of the observations is due to the experimental design used. This score plot accounts for 71% of the explained variation. Figure 9.21: (right) PCA t_3/t_4 score plot (refined model, SOVRING). This plot represents merely 19% of the explained variation. All clusters are reasonably scattered, and we are unable to detect significantly deviating clusters.

In summary, the refined PCA modeling indicated no need for further elimination of observations. The PLS modeling will, therefore, be based on the selected sub-set of 85 representative observations.

9.6.3 PLS modeling of 12 factors, 6 expanded terms, and 6 responses

Prior to using PLS, the X-matrix was expanded by a set of six additional model terms. This is possible since a composite design was developed in the three factors load (Ton_In), speed of magnetic separator 1 (HS_1), and speed of magnetic separator 2 (HS_2). The following six X-variables were added: Ton_In^2, HS_1^2, HS_2^2, Ton_In*HS_1, Ton_In*HS_2, and HS_1*HS_2.

The PLS modeling yielded a six-component model explaining 84% ($R^2Y = 0.84$) and predicting 81% ($Q^2Y = 0.81$) of the variation in the six responses (Figure 9.22). Most of the six output variables are well explained and predicted by the model (Figure 9.23). The situation is worst with regard to %P_FAR which has $R^2Y = 0.63$ and $Q^2Y = 0.56$.

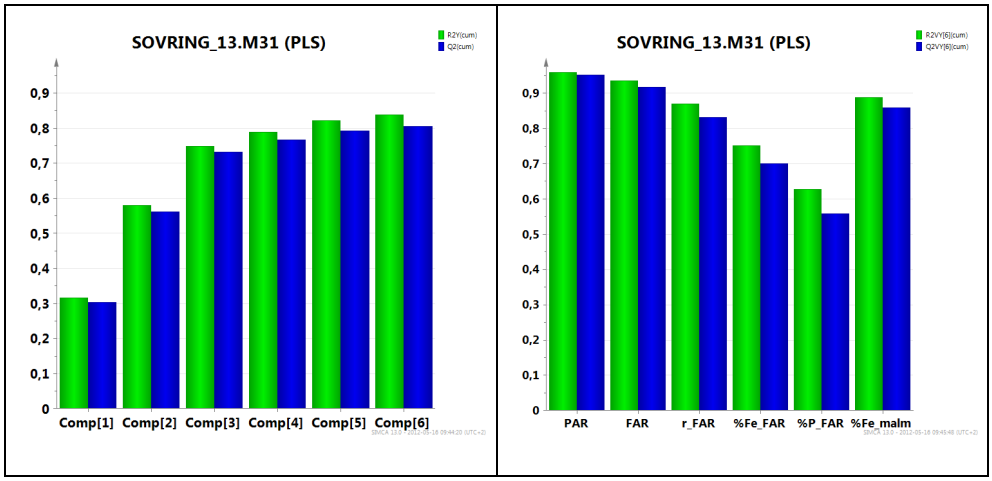


Figure 9.22: (left) Model overview of PLS model for 85 selected observations. After six components $R^2Y = 0.84$ and $Q^2Y = 0.81$. Figure 9.23: (right) Individual R^2Y- and Q^2Y-values of the six responses. The responses are amount of concentrate type 1 (FAR), amount of concentrate type 2 (PAR), distribution of type 1 and type 2 (r-FAR), iron in FAR (%Fe_FAR), phosphorous in FAR (%P_FAR), and iron in raw ore (%Fe_malm).

9.6.4 Evaluation of PLS model

In the evaluation of the PLS model, it is informative to examine the appearance of the inner relation. Figure 9.24 shows the agreement between the t- and u-scores of the first model dimension. This is a linear relationship. The relationships between further pairs of scores, t_2/u_2, to t_6/u_6, may be investigated in the same manner, but those plots are not included here. As expected, the correlation band widens for each new component, and is quite broad in a plot of t_6 and u_6. However, the last component only accounts for 2% of the response variation, so it should not be over-interpreted.

Further evaluation of the PLS model might involve consideration of DModX (Figure 9.25). The somewhat periodic appearance of this plot arises because a composite design was used in the planning of the experiments. The important aspect is that this test reveals no moderate outliers.

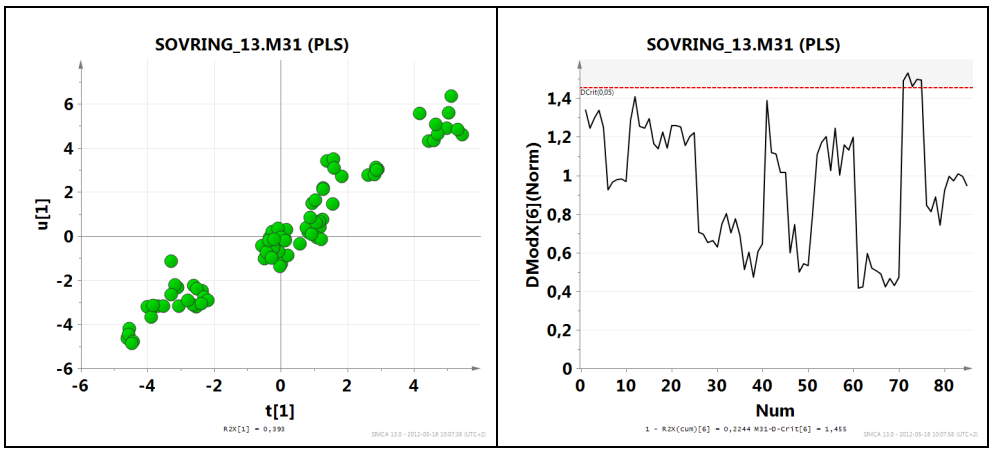


Figure 9.24: (left) PLS t_1/u_1 score plot of SOVRING model. Figure 9.25: (right) DModX plot of SOVRING PLS model. The somewhat periodic behavior arises from the underlying experimental design.

9.6.5 Interpretation of PLS model

In this section, the aim is to interpret the PLS model. For this purpose, Figures 9.26 (t_1/t_2 score plot) and 9.27 ($w*c_1/w*c_2$ weight plot) were created, which display model parameters of the first two components. One may state that the t_1/t_2 score plot looks as expected, with a uniform, but partially clustered spread of observations (compare with Figure 9.20).

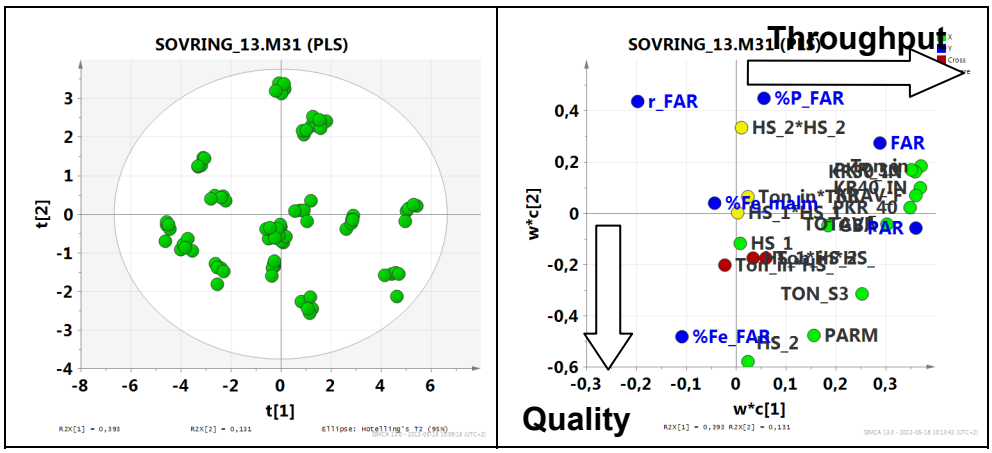


Figure 9.26: (left) PLS t_1/t_2 score plot of SOVRING model. Observations in the lower right-hand quadrant provide the best combination of high product throughput and high quality. Figure 9.27: (right) PLS $w*c_1/w*c_2$ weight plot of SOVRING model. The first component models throughput, and the second component quality of product.

By examining the corresponding weight plot (Figure 9.27), we may assign meaningful attributes to the two components. Because the variable load (Ton_In), and several other variables correlated to load, display large positive weights in the first dimension, we understand that the horizontal direction of Figure 9.26 discriminates between the observations on the basis of amount of incoming material from the mine (throughput). Samples in the right-hand part have high load, which, in turn, corresponds to high throughput of FAR and PAR.

The second component reflects quality of PAR and FAR. This is inferred from the dominant contributions of %Fe_FAR (which should be high) and %P_FAR (which should be low). Hence, the vertical direction in Figure 9.26 separates observations according to the quality of PAR and FAR produced. Observations in the lower part represent the highest quality. In summary, this means that samples in the fourth quadrant of Figure 9.26 optimally combine high load (much product) and high quality. However, we should be careful not to conclude too much from Figures 9.26 and 9.27, since we have not yet considered the higher-order components in the interpretation.

It is also worth noting the relatively low modeling power of the expanded terms. Except for HS_2^2, these terms have near-zero weight in both components. This indicates that the throughput-related responses are linearly related to the designed factors, but that quality of product has a partial quadratic dependence on the speed of the second magnetic separator.

Another way of overviewing the importance of the different X-variables is to create a VIP-plot (Figure 9.28). VIP is especially meaningful when the model has many components (here: six) and covers many responses (here: six). Among the designed factors, the speed of the second magnetic separator is most important. (Recall that because VIP is a squared function of the PLS weights, it only gives information in an absolute sense; that is, whether an X-variable is important (yes/no) but not how (positive sign/negative sign)).

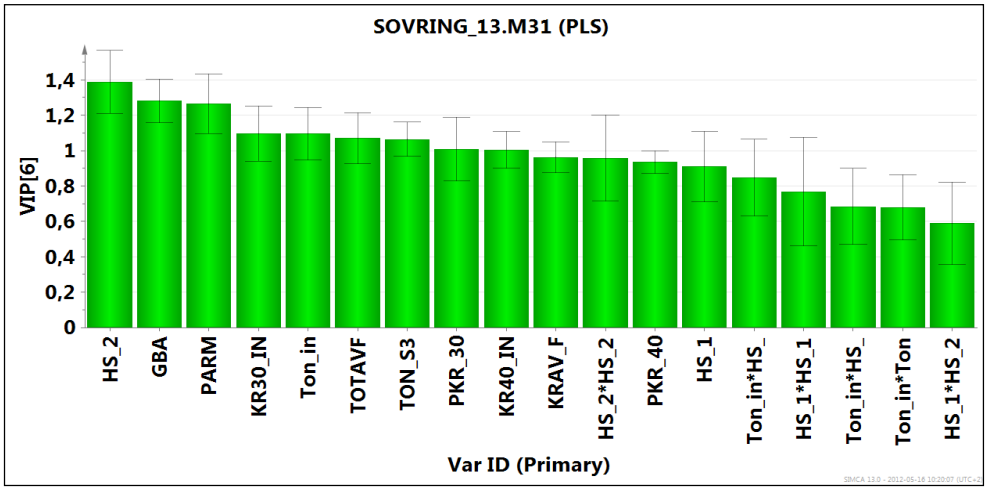


Figure 9.28: Plot of VIP of SOVRING PLS model. This type of plot is particularly attractive in situations involving many model components and many responses. Overall, the speed of the second magnetic separator is the variable with the greatest importance on the projection.

9.6.6 Response contour plotting

Since there are three X-variables, which are varied according to a composite design, we may also attempt response contour plotting in order to explore their relevance for the different responses. However, a correct contour plot cannot be established with the current PLS model, as it incorporates non-designed factors. A proper response contour plot can only be created for a model where all factors have been manipulated according to DOE.

Consequently, a modified PLS model was established in the three designed factors (Ton_In, HS_1, HS_2) plus their six expanded terms (see details in Section 18.3). This model showed performance statistics similar to those of the foregoing model, except that the last response (% iron in raw ore) was much more poorly modeled. The response % iron in raw ore cannot be influenced by the designed factors, as this quantity is determined by the natural variation down in the mine.

We will not interpret this modified PLS model in depth (see instead details in Section 18.3). For the moment, our aim is to illustrate the principle of response contour plotting, which is a highly informative way of looking at data. This approach can be used when factors have been manipulated according to an experimental design. We decided to use %P_FAR as an example response, because it was expected to show a quadratic dependence on HS_2 (cf. Figure 9.27).

The coefficient plot shown in Figure 9.29 shows that the speed of the first magnetic separator (HS_1) has limited impact on %P_FAR. In order to decrease the amount of phosphorous in the FAR concentrate, one should focus on running on a low load and use high speed of the second magnetic separator. By converting this model information into a response contour plot (Figure 9.30), we get a simpler picture to interpret. Depending on how much %P is acceptable in FAR, we realize that for each level of Ton_In there exists an optimal value of HS_2. More details regarding the use of response contour plots are found in Chapter 18.

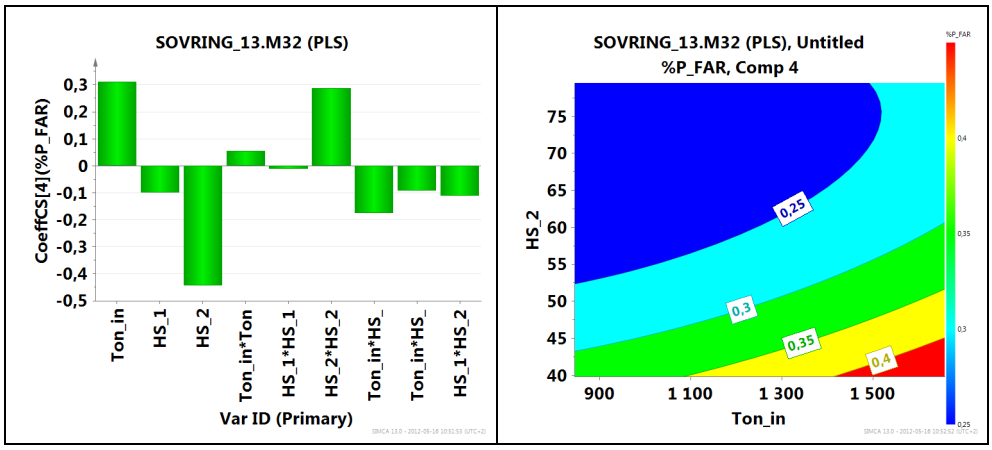


Figure 9.29: (left) Regression coefficients of %P_FAR. The PLS model is fitted to the designed factors only. The factors Ton_In and HS_2 have most impact on the response. There is significant quadratic behavior with respect to HS_2. Figure 9.30: (right) Response contour plot of %P_FAR, where the influences of Ton_In and HS_2 are seen. HS_1 has been fixed at its center level. For each level of Ton_In there exists an optimal level of HS_2 producing FAR of appropriate P-content.

9.6.7 Using the full PLS model for process monitoring

In order to monitor the process and predict responses, we must use the full PLS model based on all available descriptors. The reasons are several. On-line measurements enable on-line predictions, and using the correlation structure among all variables will aid in stabilizing the predictions. In addition, for early fault detection, it is generally advantageous to employ many descriptors rather than few, as more types of process upsets and problems can then be discovered.

The full PLS model was applied to the remaining 487 observations. The projection of these observations showed that the model immediately detected the extreme no feed situations using either Hotelling's T^2 (no plot), the t_1/t_2-plane (Figure 9.31), or DModX (Figure 9.32). There are two time periods with no feed, the first lasting from the start of the sampling period and until observation 56, and the second ranging between time points 208 and 228.

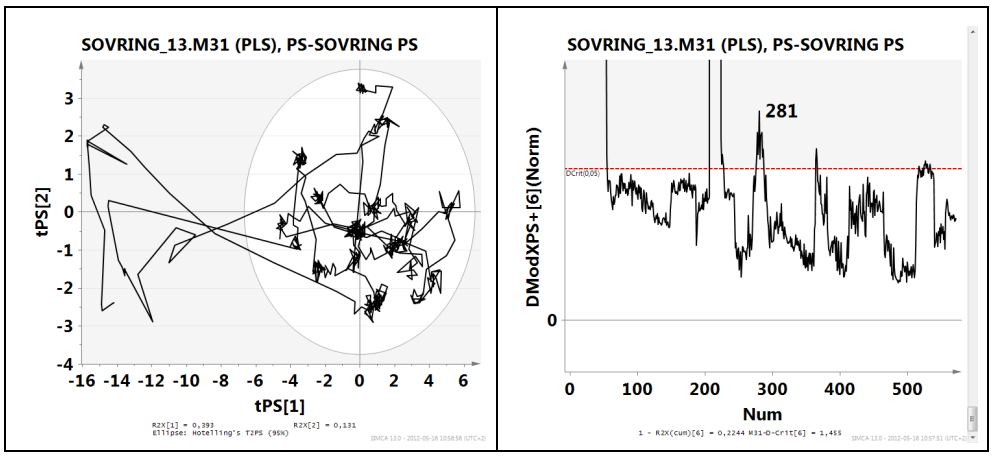


Figure 9.31: (left) PLS t_1/t_2 score plot obtained when projecting the 572 samples onto the full PLS model. The two periods of no feed, samples 1-56 and 208-228, are detected as deviations to the left of the normal operating area. Figure 9.32: (right) DModX plot obtained when fitting the 572 samples to the full PLS model. Apart from low feed observations, some others, like observation 281, have high residuals. Such samples should be inspected more closely, as predictions of Y-data for these correspond to extrapolation outside the validity range of the PLS model.

Apart from time periods of low feed, the DModX graph is also able to identify other observations with high residuals. One example that is highlighted is sample 281. The problem associated with this observation could be probed by a contribution plot (however,

no plot is shown). With regard to sample 281, many variables contribute to the abnormal residual, indicating a new short-term operating principle, rather than a sensor failure.

Sample 281 and others with high residuals must be examined carefully. It is crucial to realize that predictions for high-residual observations correspond to extrapolation outside the range of validity of the model. Estimates derived through extrapolation are less reliable than interpolations. Hence, when computing predictions, it is important to know whether the observations fit, or do not fit, the model.

9.6.8 Using the full PLS model for on-line predictions

In the mineral sorting industry, it is of particular interest to explore the possibilities of on-line predictions of iron content in the incoming ore (%Fe_malm, last response in data table). First of all, one must recognize the enormous difficulties in getting representative samples for laboratory analysis to determine the percentage of iron. Secondly, when the percentage of iron in the incoming ore becomes too low, the process people on-the-ground must inform the miners. Hence, the need for on-line predictions of this response is easily recognized.

Figure 9.33 shows a time series plot of the predicted level of iron in the raw ore. The small horizontal lines denote the measured values of the calibration samples. Underneath the time series plot, we have inserted a plot of DModX. A DModX above Dcrit indicates regions in time where on-line predictions are less reliable. Considering the nature of the response %Fe_malm, the conclusion of the original authors was that the PLS model enabled on-line predictions of sufficient quality. An indication of the predictive power can be obtained from the cross-validation parameter Q^2, which exceeds 0.8.

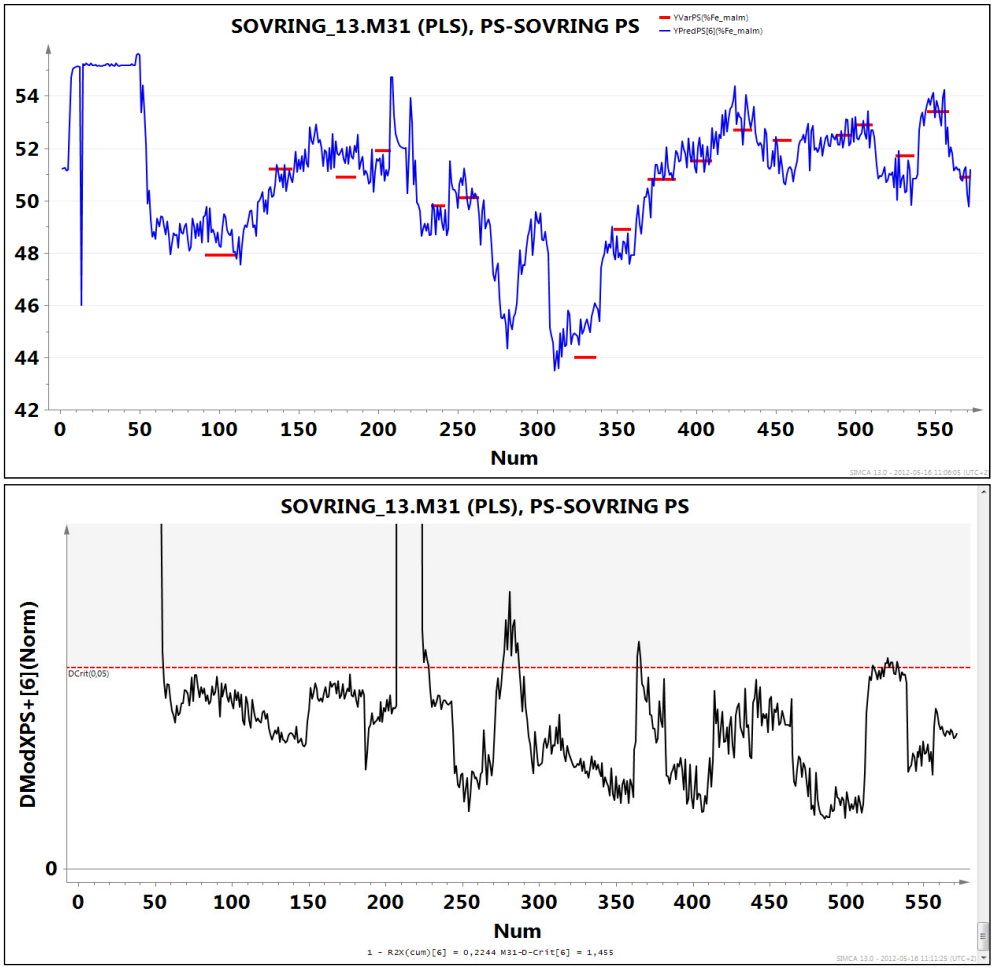


Figure 9.33: Simulation of on-line prediction of percentage of iron in incoming ore. In the upper part, the short horizontal lines represent the measured data. In the lower part of the plot, observations for which DModX exceed Dcrit should be regarded with some caution.

9.6.9 Summary of SOVRING application

This example shows that statistical experimental design in the dominating process variables gives data with high quality that can be used to develop good predictive process models. By initiating the data analysis with PCA, it was possible to discover two periods corresponding to no or low feed of starting material. These were excluded from the PLS analysis. The PLS model based on the set of 85 representative samples, enabled excellent predictions of PAR and FAR. The prediction quality was a little lower for the other responses. However, prediction of iron content in incoming ore can be accomplished with reasonable precision.

9.7 Questions for Chapter 9

1. What are the objectives of multivariate process modeling?
2. What are the two main types of process?
3. Why is it better to use a PC-score as a quality indicator rather than a univariate quality index?
4. Which tools are available to investigate normal process operating conditions?
5. Which diagnostic tool is useful for overviewing the relationships among variables?
6. Which diagnostic tool is useful for providing detailed information about a particular process change?
7. Contribution plots are available in two versions. Which?
8. How do you model normal process behavior?
9. When is it possible to use response contour plots?

9.8 Summary and discussion

This chapter has overviewed multivariate process modeling. As should be evident from the examples used, to perform multivariate process modeling properly, one must utilize not just the final product quality variables (Y) but rather all the data on process variables (X) that are routinely collected by process computers. Multivariate methods are informative and transparent, and provide powerful diagnostic tools for detecting out-of-control situations and diagnosing assignable causes. The methods are applicable to both continuous and batch processes [Kourti, 2002; Skoglund, *et al.*, 2004]. The only prerequisite for multivariate process modeling is the existence of ample historical data representing good past process performance.

We will now briefly summarize some of the available diagnostic tools. One important tool is the PCA score plot, with which it is possible to overview a process. Consider Figure 9.3, which shows an overview of the CUPRUM dataset. Copper samples lying outside the elliptic tolerance area can be considered as strong outliers. They have high leverage on the model, i.e., strong "power" to pull the PCA model toward themselves. The normal operating area in the plot is indicated by Hotelling's T^2, with, for instance, the 95% tolerance region. It must be stressed that out of 730 samples it is to be expected that around 36 samples (0.05*730) will be outside the 95% tolerance region. Only a handful of these "potential" outliers will be "real" outliers, and hence process insight by an experienced operator is needed to identify the "real" ones. Thus, one has to differentiate between outliers identified from a statistical perspective and those with practical relevance. This is simplified by the simultaneous display of 95, 99 and 99.9% tolerance levels.

In addition to scores and T^2, it is informative to look at the model residuals, usually displayed in a DModX chart. Even moderate outliers (for example: copper samples) in the data often exceed the critical distance in a DModX chart. Such outliers do not, however, have the same profound effect on the model building as do strong outliers in score space. But for process diagnostics it is important to identify outliers in residuals, because they indicate temporary process upsets of a "new" type. Persistently high DModX values indicate that a new type of longer-lasting upset or shift is taking place.

A more powerful graphical tool for process analysis is created when a scores or T^2 control chart and a DModX control chart are tiled horizontally in the same picture. This yields a chart in which both the systematic (scores, T^2) and unsystematic (DModX) variation in the process data are displayed simultaneously. Hence, one may find outliers which conform with the model and abnormal samples which break the general correlation structure (cf. Figure 9.7). The use of such combined control charts is discussed more thoroughly in Chapters 15 and 16.

Moreover, by using loading and weight plots, it is possible to get an overall understanding of process variables, and of the correlation structure between them. However, by using contribution plots, either in scores, in DModX, or both, it is possible to focus the interpretation towards a particular part of the process data, and to attribute probable causes to the observed process shifts. These tools are thus of great importance for the process analyst. The use of contribution plotting is also spreading to other types of data analytical problem.

The use of DOE is strongly recommended in multivariate process modeling. As exemplified by the SOVRING example, DOE in the dominant process factors enabled the establishment of causal relationships for an observed change, rather than just correlative indications.

Another interpretation problem with respect to process data is the presence of feed-back controllers. Feed-back controllers often influence the process such that in the modeling one obtains model coefficients of inverse sign with regard to an anticipated correlation structure. Thus, it appears to be the impact of the feedback controller that is modeled, rather than the changing properties of the process itself. This is illustrated in Figure 9.34.

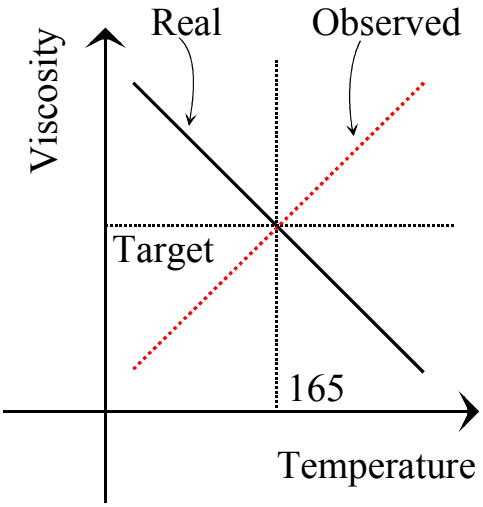


Figure 9.34: Illustration of problem with feed-back controllers. In a manufacturing application too high a viscosity was corrected by raising the temperature. The observed relationship was a positive correlation between viscosity and temperature (dashed line). The real cause and effect relationship was an inverse relationship between viscosity and temperature (solid line). When feed-back controllers are present in a process, e.g., for regulating the viscosity, the observed relationship will have the opposite sign compared with the real relationship. This happens both with automated and manual (operator) feed-back systems.

10 Classification and discrimination

10.1 Objective

In Chapter 10 we shall return to the second main type of data analytical question, *classification and discrimination* (see introduction in Chapter 1, Section 4). Classification is often done to predict a probable class membership for an observation. Discrimination is often done to sharpen the separation between groups of observations, and to understand which variables carry the class separating information. Two examples, called IRIS and ARCHAE, will be used to illustrate the principles of classification and discrimination.

10.2 Introduction

It has been said that classification is the key to knowledge about complicated systems. Man has always made abstract divisions of the observations around him into classes. Biological "observations" are divided into plants and animals, and then further subdivided into orders, families and species, e.g., dandelions and cats. Chemical compounds are classified as inorganic or organic and the latter sub-classified according to their functional groups, e.g., alcohols, amines, ketones, and hydrocarbons. Chemical reactions are classified according to type of substrate and mechanism. Solvents are classified as protic and non-protic, polar and non-polar, etc. Process time points are classified as representing good or bad process behavior, and so on.

Experience shows that nature often seems to organize itself in a clustered, rather discontinuous way. Inside a class (a cluster) the observations are rather similar to each other, so that once we know the class membership of an observation we know a great deal about it. The similarity between observations within each class is considerably greater than between observations of different classes. This is the basis for the principle of *analogy*. Once we know that a compound is a hydrocarbon, for instance, we can confidently predict how the compound reacts or fails to react with various "reagents", because we know from experience that almost all hydrocarbons behave similarly, analogously, when subjected to various "chemical treatments".

It is therefore often practical to formulate a chemical problem in terms of similarities and classes. One tries to find a battery of easily accessible properties (variables) which can be used to predict the class of an observation. One then infers that all observations within a class behave similarly with respect to other, less tractable properties. Thus the classification of an observation corresponds to the prediction of these properties. With the proliferation of quantitative data, it becomes desirable to have methods for analyzing these in terms of similarity and class. We wish to find out if the data indeed indicate that the classes are separated and, if so, how they differ with respect to each other.

10.2.1 Some typical classification situations

Depending on the problem, a number of typical classification situations may arise, in which similarity is utilized in different ways.

1. *No classes are initially defined.*
   One has measured multivariate data on a number of observations and wishes to know if the observations are strongly clustered with respect to the data. In this context, it is common to commence the data analysis with a PCA for overview of the dataset.

2. *One class is initially defined*.
   One may ask what is the typical data pattern of the class, and if this can be used to classify observations in the test set as similar to the defined class or not.
3. *Several classes are initially defined*.
   One may ask if given multivariate data can be used to distinguish between the classes and to assign test set observations to the most probable classes. In chemistry this is often called the pattern recognition (PARC) problem, while statisticians prefer the name discriminant analysis, and others call it classification.

Usually, chemists and technicians are interested in much more than a mere classification of observations. In chemical investigations it is often of interest to know how a typical class member looks in terms of the given variables, which variables are useful in the classification, how the variables relate to each other, and if there are indications for outliers or strong subgroups in some classes.

10.2.2 Approaches to class specification

There are several ways to specify a class or the similarity within a class. Below, we present them as A – D. They correspond to different degrees of knowledge and the goal is to go from less knowledge to more.

A: The first way to specify a class occurs when we know very little, and have only one property that defines a class. In organic chemistry, for instance, solvents may be classified as either polar or non-polar. In QSAR, a particular type of biological activity, say, carcinogenicity, or a curing effect on acne, often defines one class while the absence of this property defines another class.

B: A second way of specifying classes arises when we know more and have class representatives, i.e., observations that have the properties of the classes. It is not trivial to move from level A to level B. In many cases one single property in itself does not define a homogeneous class. In QSAR, for instance, carcinogenicity may be the result of many different mechanisms. Compounds which are carcinogenic by different mechanisms are not likely to constitute a homogeneous class that can be successfully modeled. The same situation is often found in other structure-property relationships. Chemical compounds which are catalysts in a reaction may, for instance, catalyze the reaction in different ways, and might therefore not be similar. Moving from knowledge level A to B also involves the problem of selecting representative observations that cover the "class domain" as well as possible. Since the latter is usually not well specified, this is a difficult problem that can only be solved in an iterative, step-wise manner.

C: On the third level of knowledge, we have information about which variables are relevant for the given problem. This information is usually obtained by a long and difficult process of trial and error, but with good chemical knowledge one can shorten this process considerably. Also, the tools of multivariate data analysis are helpful here; one important outcome of MVDA is knowledge of which variables contain the relevant information about a classification problem. Hence, one can try batteries of sensible variables to see if they work. One must, however, remember the problem of chance correlations. There is always a risk that spurious differences between classes will be found, and this risk is multiplied by the number of different ways one tries to separate the classes.

D: On the highest knowledge level, we specify each class of observations in terms of relevant variables and quantitative relations between them. This knowledge allows the classification of new observations as similar or non-similar to the class. This is the highest level of empirical knowledge, but perhaps still rather low with respect to fundamental knowledge. The relations between the variables can, however, be interpreted and utilized to increase the fundamental knowledge about the observations under investigation.

10.3 Pattern recognition (PARC)

Pattern recognition (PARC) is often described as a procedure for formulating rules of classification [Albano, *et al.*, 1978; Jellum, *et al.*, 1981]. Based on a set of given classes,

each of which contains a number of observations mapped by a multitude of variables, guidelines and rules are developed that make it possible to classify new observations as similar or dissimilar to the members of the existing classes. Thus, in some respects, pattern recognition is closely linked to the various ways of specifying a class (discussed in Section 10.2.2). Introductions to chemical and technological applications of pattern recognition can be found in the literature [Strouf, 1986; Massart, *et al.*, 1988].

PARC has been used in a wide variety of applications, only a few of which can be mentioned here. In analytical chemistry, PARC is used to classify observations like mixtures of chemical compounds, or pure compounds for which the structure is to be determined. In environmental chemistry, PARC is useful for categorizing classes of partly related pesticides, in order to understand their similarity and predict important properties, like toxicity, biodegradation, and soil sorption. In food science, PARC has been used extensively for classifying food products according to principles of geographic origin, textural properties, or physical and chemical properties. An example of the latter is found in the book by Massart *et al.* [Massart, *et al.*, 1988], where some 500 olive oils from nine Italian districts were classified on the basis of their content of eight fatty acids. Also, a growing use of PARC is seen in quality control [Andre, 2003], process monitoring [Prats-Montalban, *et al.*, 2006], food authentication [Olivieri, *et al.*, 2011] and "omics" data analysis [Musumarra, *et al.*, 2004].

Applications of pattern recognition in chemistry and other fields can be categorized into four types of problem depending on the level and scope of the study. We will now review these four levels of pattern recognition.

10.3.1 Level 1: Classification into one of a number of specified classes

On the lowest level of PARC, it is assumed that all observations in the training set and the prediction set belong to one of the initially specified classes (Figure 10.1). The scope of the analysis is then to employ the descriptor variables to achieve a classification of an unknown observation as belonging to one of the existing classes.

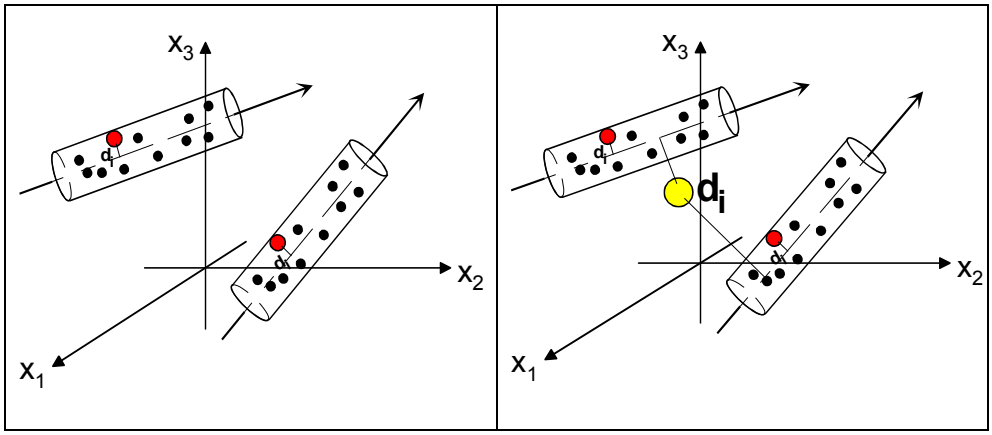


Figure 10.1: (left) PARC level 1. Observations are classified into one of several existing classes. No other classes or outliers are considered. Figure 10.2: (right) PARC level 2. As level 1, but with the possibility of outliers. The large yellow point in the middle of the graph symbolizes an observation that does not fit either of the specified classes. The fit to each class is assessed from the DModX (d_i) of the observation.

10.3.2 Level 2: Level 1 plus the possibility of outliers

On the first level of PARC it is assumed that all observations belong to one of the specified classes. This is often an unrealistic assumption, as observations in the training set might have been class-assigned incorrectly, or for some other reason be atypical for the class. Moreover, an unknown observation in the prediction set might be of a new type and mark the start of an entirely different class not previously encountered.

Thus, on level two, PARC works by representing each class of observations inside a closed mathematical structure (e.g., a cylinder, a box, an "envelope", ...) in multivariate space. This structure represents a tolerance volume, by means of which an assessment of the class membership of new observations is possible (Figure 10.2). An observation that is classified inside the tolerance volume is considered similar to the training set observations. Otherwise it is considered dissimilar.

An important special case of level 2 is called the *asymmetric case* (Level 2A), which occurs when one class has no detectable structure and the observations are spread more or less randomly. However, provided that other classes exist which have a "tight" structure, the data analysis can still be solved. Then all observations can be classified according to whether they fall inside or outside the envelope of the "tight" class(es). This is illustrated in Figure 10.3.

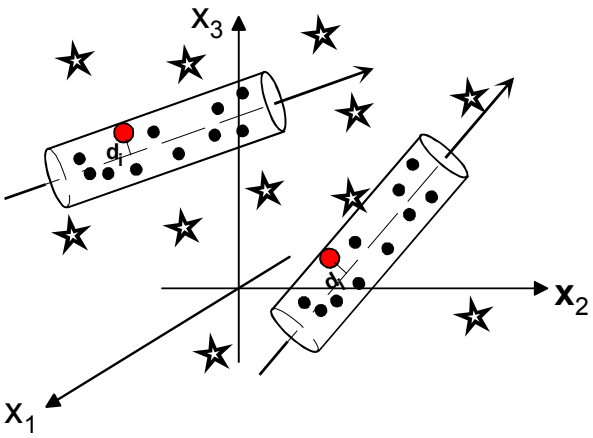


Figure 10.3: An illustration of the asymmetric case of PARC. Here, two "tight" classes form well-defined tolerance volumes, which can be used to elucidate the class membership of new observations. A new class, marked by stars, is scattered more or less randomly in the X-space. The classification of its observations can still be made, however, by observing whether the stars fall inside or outside the tolerance volumes of the "tight" classes.

A typical illustration of the asymmetric case is the active/inactive question in drug design. Usually, only a handful of chemicals are active in a set of pharmacological tests, and the structural requirements for this activity are well specified by nature. Such a group of active compounds will be clustered tightly in X-space, provided that X-variables expressing this similarity have been collected. In contrast, the kind of structural properties which make a drug inactive in the same test systems are often more diffuse and only vaguely understood. Such a class of inactives often displays a loose, semi-random scatter in the X-space.

Another example of the asymmetric case is often encountered in process modeling, where the "tight" class is made up of process time points representing good or acceptable process performance, and the "scattered" class by observations encoding undesired process conditions.

10.3.3 Level 3: Level 2 plus a response variable (y)

The third level of PARC involves the modeling of a response variable with PLS. Here, the objective of the modeling becomes (i) to devise a classification of observations on the basis of their X-variables, and (ii) to predict the numerical values of the response variable. As outlined in Section 4.4.1, the objective of the PLS analysis is then to find out whether a relationship exists between the way the observations are grouped in the X-space and the way they spread in the response space. If a new observation is inside the tolerance volume of a class, the predicted value of the response variable is computed. If a new observation is located outside the tolerance volume, the prediction may still be computed but should be viewed as less reliable.

10.3.4 Level 4: Level 3, but with a matrix of responses (Y)

Level four is probably the most realistic level of PARC. On this level we have the same X-data as on level three, but the response data have been determined in several test systems, leading to a matrix of responses (Y). The level four problem might be easier to handle than the level three problem, since the presence of several correlated response variables stabilizes the systematic structure in the data, thereby stabilizing the PLS model. The data analysis includes the two phases described above for level three [(i) classification, (ii) prediction of Y].

10.4 Example: IRIS

As an illustration of the various steps of classification, we shall now consider the IRIS example. We recall from Section 1.6.1 that this dataset includes the length and width of sepals and petals of three species of Iris, i.e., *Iris Setosa*, *Iris Versicolor*, and *Iris Virginica* (Figure 10.4). For each species, 50 observations were compiled, of which 25 were used as a training set and 25 as a prediction set. Thus, the X-matrix of the combined training set consisted of 75 rows (25 Setosa, 25 Versicolor & 25 Virginica) and 4 columns. Similarly, the prediction set contained 25 Setosa, 25 Versicolor, and 25 Virginica.



Figure 10.4: Photographs of Iris Setosa (left), Iris Versicolor (middle) and Iris Virginica (right).

10.4.1 An overview PCA model

In order to overview the IRIS data, we calculated two principal components on the training set. With these we obtained an explained variation of 96% ($R^2X = 0.96$) and a predicted variation of 75% ($Q^2X= 0.75$). The scores and loadings of this model are plotted in Figures 10.5 and 10.6, respectively.

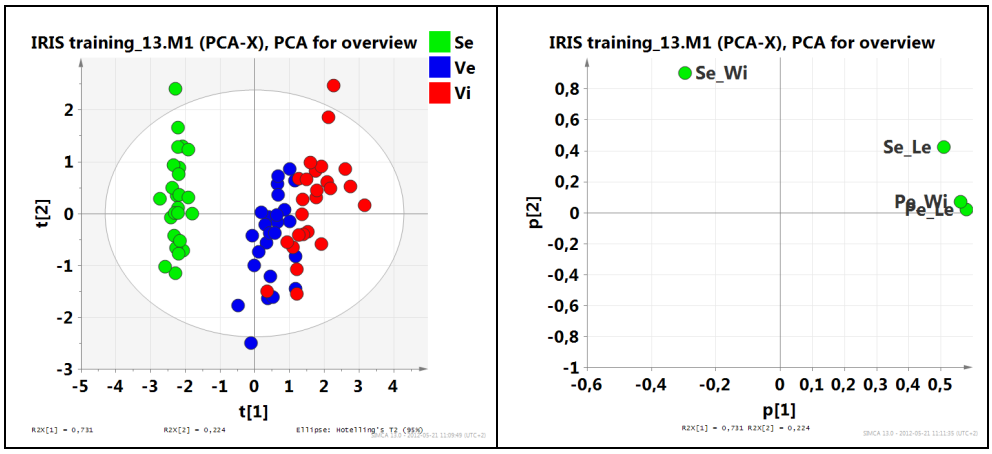


Figure 10.5: (left) PCA t_1/t_2 score plot of the overview model of the IRIS training set. The clustering of the observations shows that Setosa is markedly different from Versicolor and Virginica. It is the first component that separates Setosa from Virginica and Versicolor. Figure 10.6: (right) PCA p_1/p_2 loading plot corresponding to Figure 10.5. In the first component, width and length of petal, and length of sepal dominate. This indicates that specimens of Setosa generally have slimmer and shorter petals, and shorter sepals, than specimens of Versicolor or Virginica.

As shown by Figure 10.5, Setosa separates from the other two types of Iris flowers. This separation chiefly occurs along the first principal component. According to the loadings, it is the three variables sepal length, petal length, and petal width that encode this separation. Thus, the model interpretation suggests that the Setosa class has shorter and slimmer petals, and shorter sepals, than the other two classes.

It is not easy to see if the data contain class separating information between Virginica and Versicolor. In order to explore this, all Setosas must be eliminated from the combined training set, and a new PCA model fitted.

10.4.2 Fitting a new PCA model to the Versicolor and Virginica classes

A new PCA model was fitted to the modified training set of 50 flowers, i.e., 25 Versicolor and 25 Virginica ($R^2X = 0.87$, $Q^2X = 0.44$, A = 2). Note that the second component was not significant according to cross-validation. However, two components were computed to allow score plotting in scatter plots. Figures 10.7 and 10.8 show the scores and loadings of the refined PCA model.

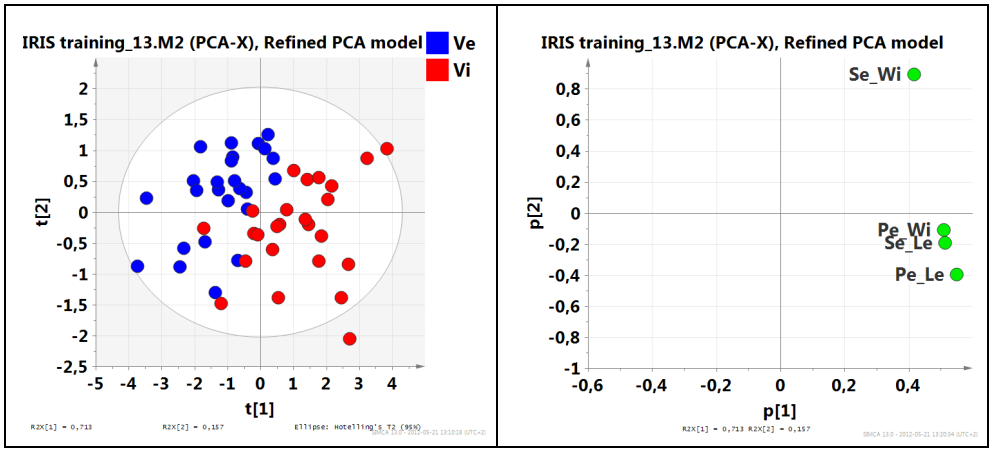


Figure 10.7: (left) PCA t_1/t_2 score plot of the refined model of the IRIS dataset. There is still a considerable overlap of Virginica and Versicolor specimens. Figure 10.8: (right) PCA p_1/p_2 loading plot corresponding to Figure 10.7. The model interpretation is similar to the foregoing model, that is, the first component indicates the sepals and petals of the Virginica specimens to be longer and wider.

According to Figure 10.7, the majority of the Virginica observations are found in the right-hand part of the score plot, and the majority of the Versicolor observations in the left-hand area. Despite the slight separation, there is still an overlap between the two species of Iris in the central part of the score plot. The loadings (Figure 10.8) indicate that specimens of Versicolor are generally smaller than members of the Virginica family.

The intriguing question is whether it is possible to further fine-tune the models such that they are able to distinguish completely between Iris Versicolor and Iris Virginica (and Iris Setosa), and, if so, would there be a nice graphical way to summarize the classification results? For this purpose, we will now consider a methodology based on disjoint principal component analysis of each Iris species.

10.4.3 Soft independent modeling of class analogy (SIMCA)

Data that are observed within a class of similar observations can always be well approximated by a few-component PCA model, provided that most of the descriptor variables indeed express this similarity. This is the basis of a method known as *soft independent modeling of class analogy*, or SIMCA for short [Wold, *et al.*, 1984].

SIMCA is a graphically oriented technique that operates on PARC level two and higher. It usually starts with a PCA on an entire dataset to get an overview of the data. Such an overview of the training set data gives indications of class separation, trends, and outliers.

The data should be selected such that each class of observations contains homogeneous data material. Provided that the dataset is not pruned inappropriately, e.g., to artificially enhance a class separation, a local PC-analysis of each resulting homogeneous class will then be useful for classifying new observations.

Subsequent to the initial data overview, in the SIMCA method each class of observations is modeled separately by disjoint PCA models (Figure 10.9). In order to unravel the appropriate dimensionality of each local PCA model, we recommend that the cross-validation technique be used. Based on the residual variation of each class, one can compute the distance to the model (DModX) of each observation. It is also possible to compute a critical distance for each class model. This was discussed in Section 3.5.1.

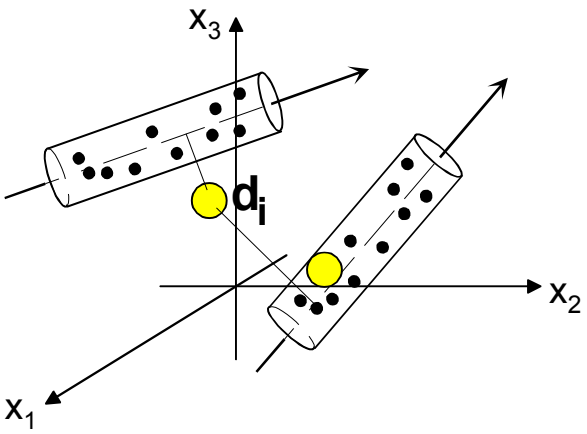


Figure 10.9: In the SIMCA methodology, disjoint principal component models are fitted to each existing, homogeneous class of similar observations. Here, the drawing applies to one-dimensional models of each class. From the scatter of points around these models, one can construct tolerance cylinders, indicating the volume spanned by each class. New test set observations are classified as class members if they fall inside a certain tolerance cylinder.

After the separate modeling of each class, the models are used to predict a likely class membership ("classification") for new observations. An observation is classified in SIMCA according to the tolerance intervals of the different classes. Observations that do not fit any class are then considered as outliers, or perhaps as founders of a new, hitherto unseen, class. Furthermore, in regions where tolerance intervals overlap, the observations cannot be unequivocally assigned.

These local PCA models can be interpreted in the same manner as outlined in Chapters 3 and 9, by inspecting loadings, scores, residuals, and contribution plots. Among other things, this will indicate which variables contribute to modeling class similarity (loading plot), and which variables do not (contribution plot in residuals of non-fitting observations).

10.4.4 SIMCA of the IRIS dataset

SIMCA was applied to the IRIS data. Thus, three local one-dimensional models were computed, each founded on 25 specimens of either Setosa, Versicolor, or Virginica. The classification performance of each model was assessed by predicting class memberships of the 75 observations (25 Setosa, 25 Versicolor & 25 Virginica) in the prediction set. The results from the classification procedure are summarized by Figures 10.10 – 10.12.

Figure 10.10 shows how the 75 prediction set observations fit the model trained on the 25 Setosa observations. The critical distance displayed corresponds to the 0.05 level, and defines a 95% tolerance interval. We find that almost all Setosa observations of the prediction set are predicted within the critical distance. In other words, a model trained on Setosa recognizes this type of flower in the classification phase. Moreover, it is evident that the other two species are far outside the tolerance interval of the Setosa model. This means that the compiled descriptor data find Versicolor and Virginica to be different from Setosa.

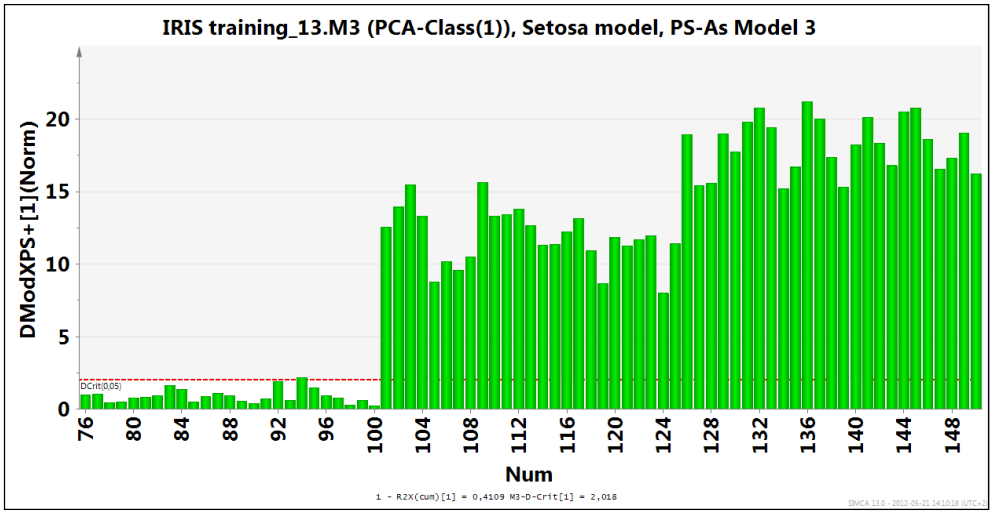


Figure 10.10: DModX plot for the IRIS prediction set of the Setosa model. Observations 76-100 belong to the Setosa family and are correctly classified, with the exception of observation 94. Observations 101-150 do not fit the Setosa model, as their DModX-values exceed the critical distance.

The classification step was repeated for the model trained on the 25 Versicolor observations, the results of which are seen in Figure 10.11. Evidently, all Versicolors (observations 101-125) of the prediction set are correctly classified. It is also clear that the Setosas (observations 76-100) are assigned as non-members. In contrast, however, the classification with regard to the Virginicas (observations 126-150) is ambiguous; some are classified as class members while others are not.

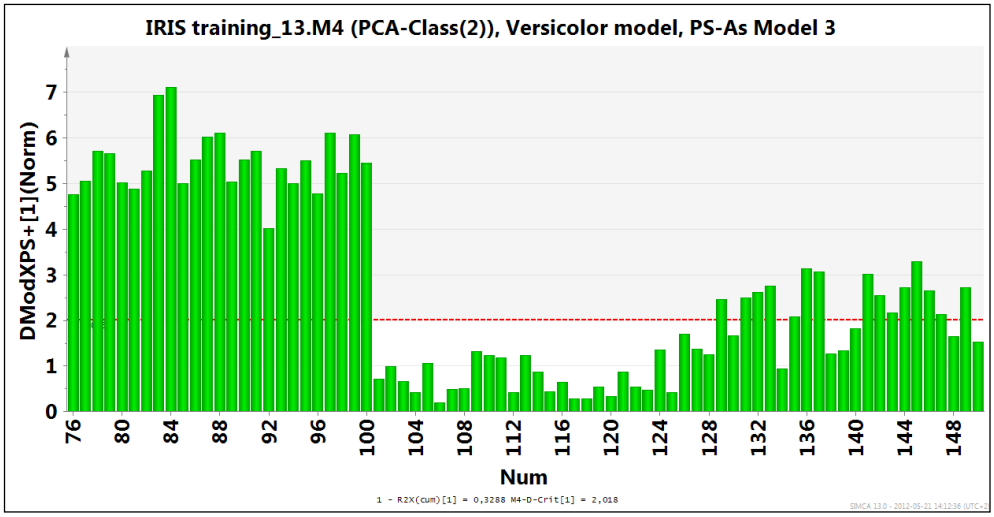


Figure 10.11: DModX plot for the IRIS prediction set of the Versicolor model. Observations 101-125 belong to the Versicolor class and are correctly classified. An ambiguity exists with regard to the Virginica samples (observations 126-150); some fit and some do not fit the Versicolor model. It is also evident that the 25 Setosa specimens (observations 76-100) are classified as very different from the Versicolor flowers.

When using the last model, the one developed for the 25 Virginica observations, results very similar to the foregoing case were obtained (Figure 10.12). Thus, the Virginicas (observations 126-150) were all correctly predicted, some Versicolors (observations 101-125) were classified as members of the Virginica class, while others were classified as non-members, but the Setosas (observations 76-100) were always found outside the tolerance interval of the Virginica model.

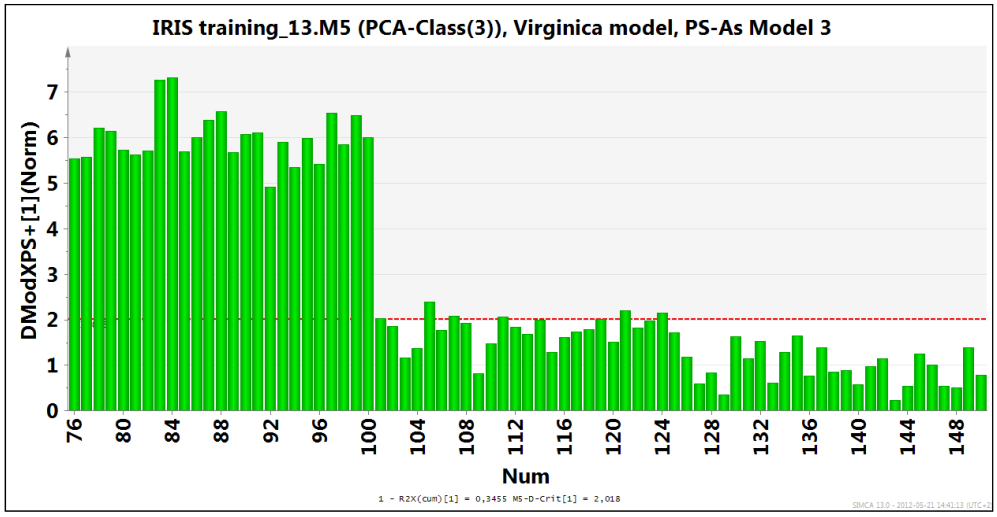


Figure 10.12: DModX plot for the IRIS prediction set of the Virginica model. Observations 126-150 are correctly assigned as Virginica samples. The classification performance is less clear-cut with respect to the Versicolor observations (observations 101-125); some fit and some do not fit the Virginica model. It is also evident that the 25 Setosa specimens (observations 76-100) are categorized as very different from the Virginica samples.

10.4.5 The Coomans' plot: A powerful graphical display

In addition to a bar chart of DModX's (cf. Figure 10.10 – 10.12) of a series of observations, results from a classification analysis may be nicely presented in a plot called the *Coomans' plot* [Coomans, *et al.*, 1984; Wold, *et al.*, 1984]. This plot is named after the Belgian chemometrician Danny Coomans, who in the mid 80´s demonstrated its great applicability. The essence of the Coomans' plot is that class distances (DModX's) for two classes are plotted against each other in a scatter plot. This is exemplified in Figure 10.13 with the IRIS data.

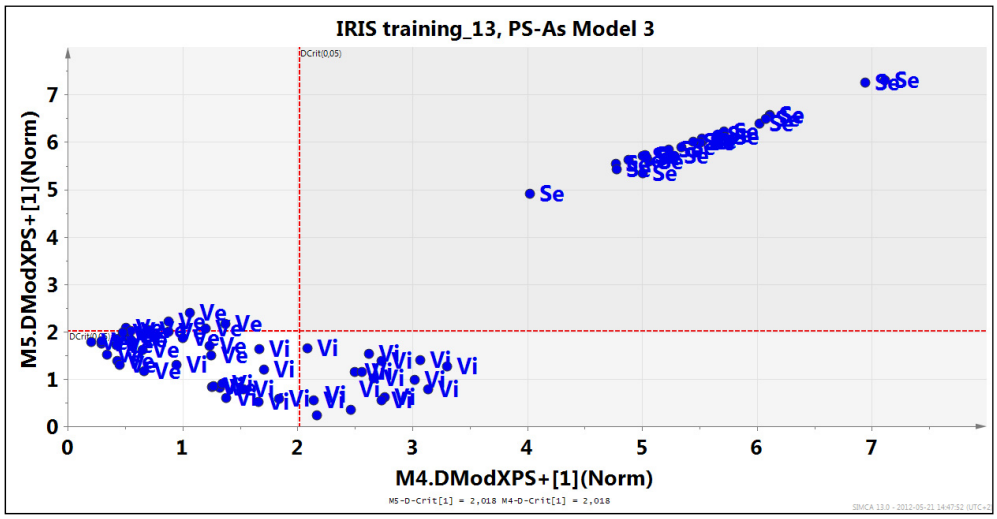


Figure 10.13: A Coomans' plot of the SIMCA classification results of the IRIS dataset. The co-ordinates are the DModX values for the prediction set observations from the Versicolor model (x-axis) and the Virginica model (y-axis). The critical distances are indicated by the solid lines, and they correspond to 95% tolerance intervals.

By plotting also the critical distance, DCrit, for each model in the Coomans' plot, four areas of diagnostic interest are created. Consider Figure 10.13, which represents a scatter plot of DModX to the Virginica model against DModX to the Versicolor model. In the lower left-hand part of the plot there is a region where prediction set samples that fit both models are found. In the lower right-hand part and the upper left-hand part there are regions where those observations predicted to fit the Virginica model or the Versicolor model are found,

respectively. Finally, we have the upper right-hand area where we find observations that do not conform with either of the models.

The conclusions that may be drawn based on the Coomans' plot are thus the following:

(i) Setosa species are quite different from Versicolor and Virginica observations.

(ii) There is an overlap between the Versicolor and Virginica classes, and they cannot be completely separated.

10.4.6 Summary of the IRIS application

An overview PCA was made on the training set of 75 observations. This model revealed that Setosa flowers were different from the other two species. In an attempt to better resolve the Versicolor and Virginica samples, a new model was founded on these two groups. Only a partial separation was obtained. Finally, the SIMCA approach was tested on the IRIS dataset. By constructing three local PCA models pertaining to each separate class of Iris species, and predicting the probable class membership of 75 prediction set observations, it was corroborated that Setosa observations were markedly different from the others. Through the construction of a Coomans' plot, it was also confirmed that the classes of Iris Versicolor and Iris Virginica overlap in multivariate space.

In Sections 10.8.4 and 10.8.5 below, the Coomans' plot and its relation to two additional classification tools, the classification and misclassification lists, is outlined using the IRIS dataset.

10.4.7 A recent extension to the IRIS dataset

A few years ago, Mark Earll (Umetrics, UK) extended the Fisher IRIS dataset by measuring equivalent data for a set of 40 flowers. Thirty-three of these flowers were specimens of Iris Pseudacorus and seven were unknown Irises. The class of Iris Pseudacorus flowers has a common name of “Yellow Flag”. Obviously, the Yellow Flags can be told apart from the Setosas, Virginicas and Versicolors by simply using the color difference (Figure 10.14). However, we may still use the data measured on these additional flowers in a tutorial perspective, to illustrate how existing multivariate models can be used on a routine basis to classify new samples as similar or dissimilar to the training set samples. This will allow us to answer questions, such as:

- How do the new flowers fit the existing PCA models?
- If they don´t fit, how do they differ from the training set?
- And what about the classification of the seven unknowns?

Below, we shall demonstrate how the existing PCA models can be interrogated in a sequential manner in order to address these questions.



Figure 10.14: Photographs of Iris Setosa (left), Iris Versicolor (middle left), Iris Virginica (middle right) and Iris Pseudacorus (right).

10.4.8 Classification of the extended dataset into the overview PCA model

The 40 flowers of the extended dataset were fitted to the initial PCA model developed from the training set. Figure 10.15 displays the scatter plot of the first two score vectors (this plot is identical to Figure 10.5) and Figure 10.16 shows the outcome of the classification. It is

immediately realized that the seven unknowns are very dissimilar to the 75 training set flowers, because they are situated outside the tolerance ellipse of the model. The Yellow Flags are closest to the Virginicas.

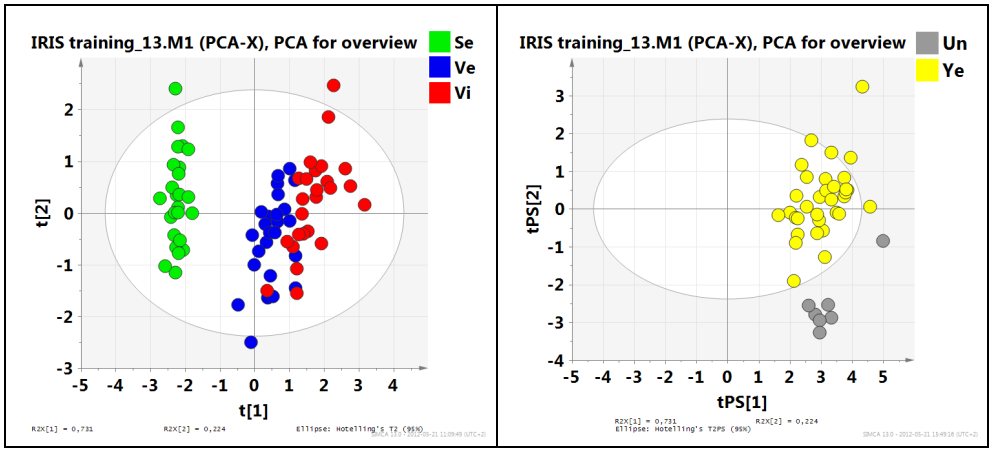


Figure 10.15: (left) PCA t_1/t_2 score plot of the overview model of the IRIS training set. Note that this plot is identical to Figure 10.5. Figure 10.16: (right) Predicted t_1/t_2 score plot showing the distribution of the Yellow Flags and the Unknowns when referenced against the distribution of the Setosas, Versicolors and Virginicas of the training set. The conclusion is that the Unknowns are very dissimilar to the training set flowers.

10.4.9 Classification of the Yellow Flags into the local Virginica model

The results in the foregoing section showed that (i) the Unknowns were very different to the training set flowers and that (ii) the Yellow Flags were most similar to the Virginicas. Hence, it is of relevance to investigate if the Virginicas and the Yellow Flags can be better resolved from one another using the local class model fitted to the Virginicas. Figure 10.17 shows a plot of the single score vector of the Virginica model. Similarly, Figure 10.18 provides a plot of the corresponding predicted score vector for the Yellow Flags. Note that the ordinate in both Figure 10.17 and 10.18 have been matched to facilitate comparison of the plots.

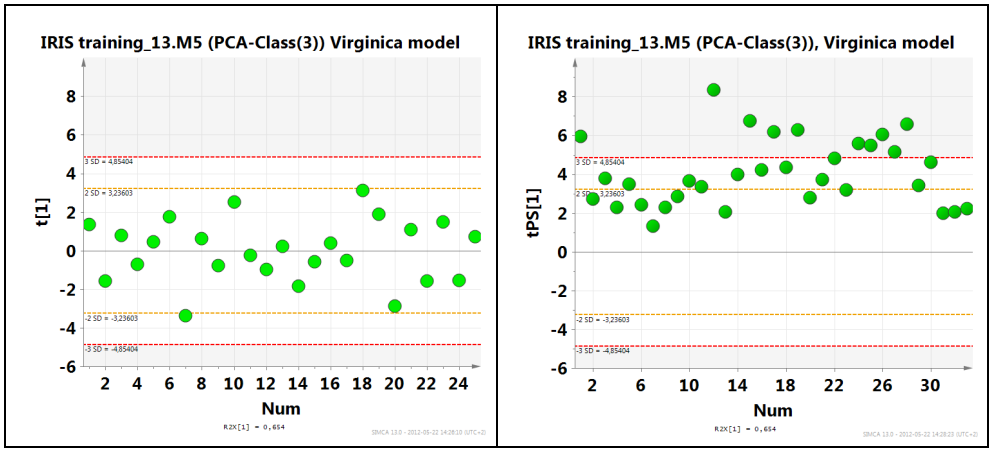


Figure 10.17: (left) Score plot for the single score vector of the Virginica model. Figure 10.18: (right) Predicted score plot for the Yellow Flags when fitted to the model established on the Virginica samples.

It is evident from Figures 10.17 and 10.18 that all Yellow Flags are positioned on the numerical high-end side of the score vector, but the two groups are not fully separated. Interestingly, however, the two groups are completely separated using the DModX statistic (Figures 10.19 and 10.20). All Yellow Flags are clearly outside the tolerance cylinder

around the Virginica model with a DModX for every flower that is at least twice the critical distance.

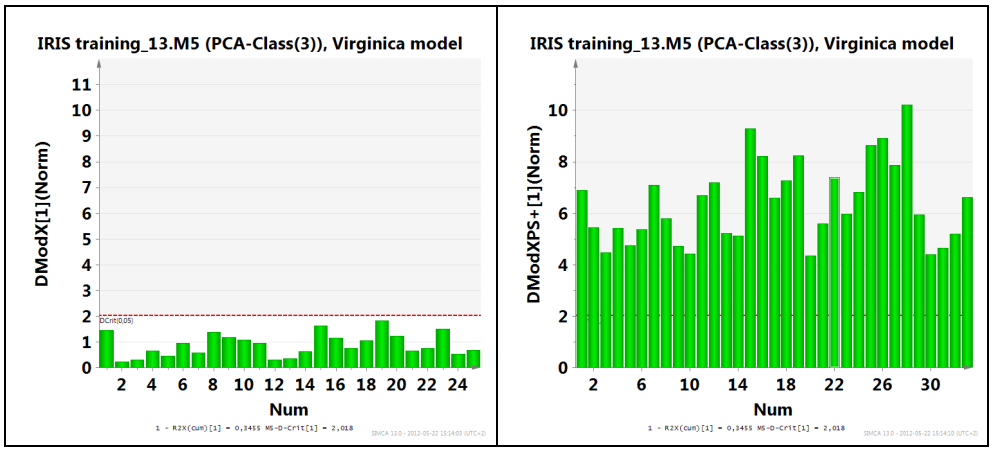


Figure 10.19: (left) DModX plot of the Virginica model. Every Virginica sample is an inlier. Figure 10.20: (right) Predicted DModX plot for the Yellow Flags obtained when fitting them to the model for the Virginica flowers. The DModX for each member of the Yellow Flag class is at least twice the critical distance to the model. This fact shows that the Yellow Flags are completely separable from the Virginicas.

10.4.10 Group contribution plot to compare two groups

The group contribution tool can be used to assess in what way the Virginicas and Yellow Flags differ. Figure 10.21 shows a group to group comparison (scores mode) that suggests the main difference be ascribed to Petal_Width; the Yellow Flags have much wider Petals. Figure 10.22 is a line plot of the original measurements observed for Petal_Width. Apparently, the average Petal Width for the Virginicas is around 2.0 cm whereas the corresponding average for the Yellow Flags is approximately 4.2 cm. Hence, the raw data plot confirms the finding of the group contribution plot.

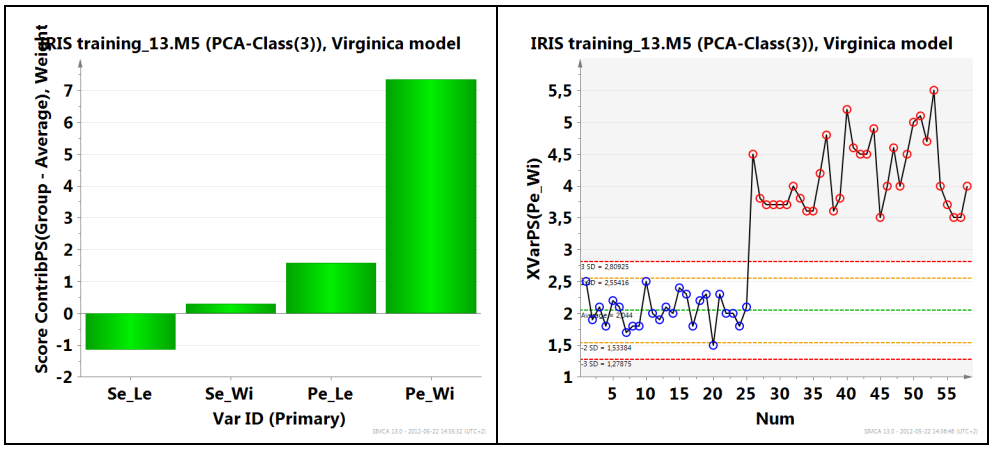


Figure 10.21: (left) Group contribution plot comparing Virginicas and Yellow Flags. The plot indicates that the main difference lies in the variable Petal_Width, for which the Yellow Flags have higher numerical values. Figure 10.22: (right) Raw data plot of the variable Petal_Width. Blue circles denote Virginicas and red circles denote Yellow Flags. On average Petal_Width is approximately 2.0 cm for Virginias and 4.2 cm for Yellow Flags.

10.4.11 Final summary of the IRIS application

In addition to what was expressed in Section 10.4.6, the extended IRIS dataset shows how the previously developed models (the overall model and the local Virginica model) can be deployed in a sequential manner. Using these models we infer that

- the Unknowns are not similar to the Setosas/Versicolors/Virginicas;

- the Yellow Flags are most similar to the Virginicas;
- the Yellow Flags are completely separable from the Virginicas using DModX;
- the main difference (Yellow Flags vs. Virginicas) lies in Petal_Width.

10.5 Partial least squares discriminant analysis (PLS-DA)

When a single PCA is used in a PARC problem with a set of observations representing one or several classes, knowledge related to class membership is *not* used to find the location of the principal components. The PCA model is calculated to approximate the observations as well as possible. It must be realized that PCA finds the directions in multivariate space that represent the largest sources of variation, the so called principal components. However, it is not necessarily the case that these maximum variation directions coincide with the maximum separation directions among the classes. Rather, it may be that other directions are more pertinent for discriminating among classes of observations.

It is in this perspective that a PLS based technique, called PLS discriminant analysis (PLS-DA), becomes interesting [Sjöström, *et al.*, 1986; Ståhle, *et al.*, 1987; Worley, *et al.*, 2013]. PLS-DA makes it possible to accomplish a rotation of the projection to give latent variables that focus on class separation ("discrimination"). The method offers a convenient way of explicitly taking into account the class membership of observations even at the problem formulation stage. Thus, the objective of PLS-DA is to find a model that separates classes of observations on the basis of their X-variables. This model is developed from a training set of observations of known class membership.

In PLS-DA, the X-matrix consists of the multivariate characterization data of the observations. In order to encode a class identity, one uses as Y-data a matrix of dummy variables, which describes the class membership of each observation in the training set. A dummy variable is an artificial variable that assumes a discrete numerical value in the class description. The dummy matrix Y has G columns (for G classes) with ones and zeros, such that the entry in the g.th column is one and the entries in other columns are zero for observations of class g (Figure 10.23).

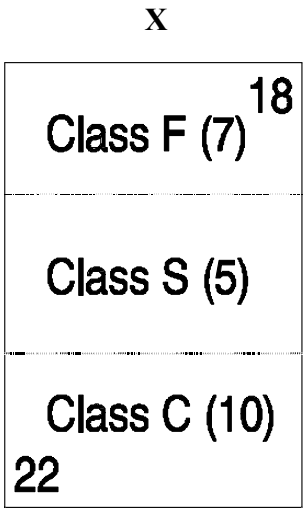


Figure 10.23: In PLS-DA, an artificial "dummy" variable is added for each category of observations. PLS is then used to relate X and Y. In the example, there are three classes of observations, denoted F, S, and C (see text). Three dummy variables are utilized to account for the class membership of each observation.

Provided that each class is "tight" and occupies a small and separate volume in X-space, one can find a plane - a discriminant plane - in which the projected observations are well separated according to class. However, when some of the classes are not tight, e.g. because of a lack of homogeneity and similarity in the non-tight classes, the discriminant analysis does *not* work. Then other approaches, such as SIMCA, must be used, where a PCA or PLS

model is developed for each tight class, and new observations are classified according to their proximity in X-space to these class models.

Typical applications of PLS-DA in chemistry include the classification of samples according to their origin in space (e.g., a French or Italian wine) or time (i.e., dating), the classification of molecules according to properties (e.g., acid or base, beta receptor agonist or antagonist), or the classification of reactions according to mechanism (e.g., S_N1 or S_N2). Typical investigations in process modeling involve the classification of process behavior as "good" or "bad". Another interesting application was published by Nouwen *et al.* [Nouwen, *et al.*, 1997], who employed PLS-DA in the classification of environmentally occurring chemicals. Based on structural fragments, they were able to classify some 500+ high production volume chemicals into classes like inert, less inert, reactive, and specifically acting compounds. And in medical research, Gottfries and colleagues used PLS-DA to diagnose different types of dementia [Gottfries, *et al.*, 1995].

Finally, we observe that classical regression analysis may also be used in discriminant modeling. However, this requires that the X-variables are few and independent (i.e., the regression assumptions are fulfilled). It is also necessary to construct a Y-matrix that has G-1 columns (for G classes) with ones and zeros, such that the g.th column is one and the others zero for observations of class g when g < G-1. For class G all columns have the value of -1.

This concludes the introductory description of PLS-DA. We will now apply SIMCA and PLS-DA to a dataset called ARCHAE.

10.6 Example: ARCHAE

One important issue in archaeology is characterization and classification of soil samples to determine their origin and age. Often this is possible since undisturbed soil and soil modified by cultivation or dwelling exhibit quite different patterns of trace element and main component composition. In many applications, mapping of phosphate has become more or less standard in the characterization of soil samples. However, this is a univariate description which does not necessarily relate perfectly to prehistoric activity. Realizing the need for multivariate characterization of soil samples using many descriptors, Johan Linderholm and coworkers at Umeå University carried out a feasibility classification study involving 18 chemical descriptors [Linderholm, *et al.*, 1994].

Linderholm and coworkers collected soil samples from an excavation area outside Mjölby in Sweden. The collected samples represent three categories: samples from recognizable features (F) from the site, samples from the occupation layer of the site (S), and off-site control samples (C). In total, 22 samples were assembled representing varying degrees and intensity of human cultural influence (see Figure 10.23). To carry out multivariate chemical characterization of the soil samples, the levels of nine elements (Fe, Cu, P, Mn, V, Co, Zn, Cr, and Ca) were recorded. These measurements were made using ICP-AES and testing two kinds of pre-treatments, viz. total dissolution (TD) and nitric acid (NA). Thus, the chemical characterization embraced 18 descriptors.

In the following, the archaeological data will be investigated by three approaches, namely (i) overall PCA modeling, (ii) SIMCA, and (iii) PLS-DA.

10.6.1 An overview PCA model

Prior to the data analysis, the 18 chemical descriptors were log-transformed in order to obtain a more normal distribution of the data. Data pre-treatment involving transformations is further addressed in Chapter 12.

The PCA modeling yielded a two-component model with $R^2X = 0.93$ and $Q^2X = 0.83$. In the t_1/t_2 score plot (Figure 10.24) the three categories of archaeological samples are well separated. The diagonal going from the upper left-hand corner to the lower right-hand corner might be interpreted as a direction of cultural activity. The corresponding loading plot (Figure 10.25) reveals that the trace elements cluster in two groups. One group contains Mn, Cu, Ca, Zn and P and the other Co, V, Fe, and Cr. There is a larger difference between

these two groups of variables than between the TD and NA chemical pre-treatment methods. Hence, the NA method, which is faster than the conventional TD procedure, may be used as a replacement for the latter.

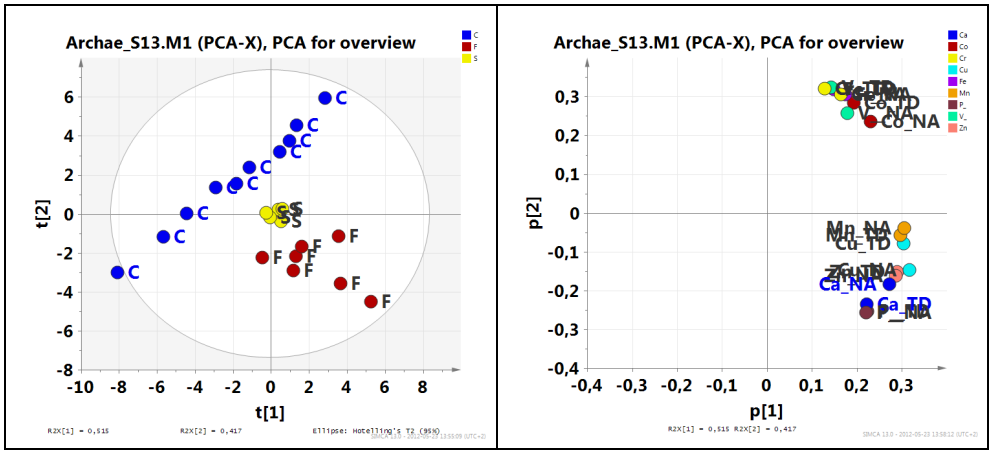


Figure 10.24: (left) PCA t_1/t_2 score plot of the overview model of the ARCHAE dataset. The observations are clustered in accordance with the pre-assigned classes. Figure 10.25: (right) PCA p_1/p_2 loading plot corresponding to Figure 10.24. Apparently, the variables cluster in two groups indicating two patterns of trace element and main component profiles. Of interest for the analytical chemist is the fact that the variation between the two sample pre-treatment methods (NA and TD) is much smaller than the variation between the different elemental variables. Hence, the NA method, which is faster than the conventional TD procedure, may be used as a replacement for the latter.

10.6.2 SIMCA of archaeology data

Since the observations were clustered according to class in Figure 10.24, we decided to further explore the inner structure within each class by means of the SIMCA approach. The aim was also to investigate whether it would be possible to use the local PCA models for classification.

Accordingly, PCA was used to model each class separately. Cross-validation was used to determine the optimal complexity of each class model. We found that two components were necessary to adequately model the seven F-samples, one component for the five S-class observations, and two components for the ten members of class C. These models define class (tolerance) volumes of different complexity. The extent of each tolerance volume is indicated by a critical distance. Figures 10.26 – 10.28 give the classification results obtained when applying each class model to the complete set of 22 observations.

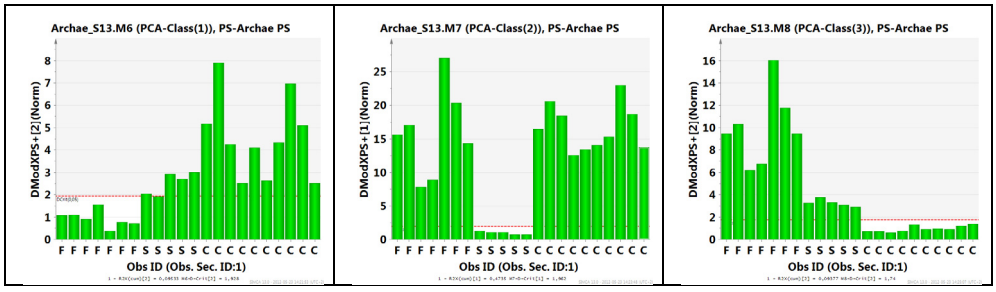


Figure 10.26: (left) Classification results of Feature model. Figure 10.27: (middle) Classification results of Site model. Figure 10.28: (right) Classification results of Control model. In each graph, the critical distance at the 0.05 level is displayed.

The classification results displayed in Figures 10.26 to 10.28 are quite different among the class models. Evidently, the Feature model is least specific (Figure 10.26), since there are a few observations (e.g. nos. 8, 9, and 18) classified as being similar to the F-class observations. Considering the shape of Figure 10.24, these findings may appear a little surprising, since the F-samples are clustered together in an area where there are no other

classes. The explanation is probably that in the local modeling of the F-class a rotation occurs, such that the C- and S-samples are projected much closer to the domain of the F-class.

Interestingly, the model of class C (Figure 10.28) is more specific than the F-model. This is because only Control samples are found inside the critical distance. According to this model, there is no doubt that Site observations have more in common with Control samples than have Feature observations. Furthermore, by looking at Figure 10.27, we find the model of class S to be most specific. All F- and C-observations have DModX-values exceeding the critical distance of the Site model by a factor four or more. This classification arises because the Site class is the most homogeneous one (see Figure 10.24 where these observations occupy a small region in score space). In part, this homogeneity is due to the low number of class representatives (five).

We also exploited the Coomans' plot option. Figure 10.29 shows a Coomans' plot where the co-ordinates are the DModX values for the prediction set observations down onto the Feature model (x-axis) and the Control model (y-axis). The critical distances are indicated by the solid lines, and they correspond to 95% tolerance intervals. The conclusion is that two Site samples share some common properties with the members of the Feature class. However, the Feature samples and Control samples do not overlap.

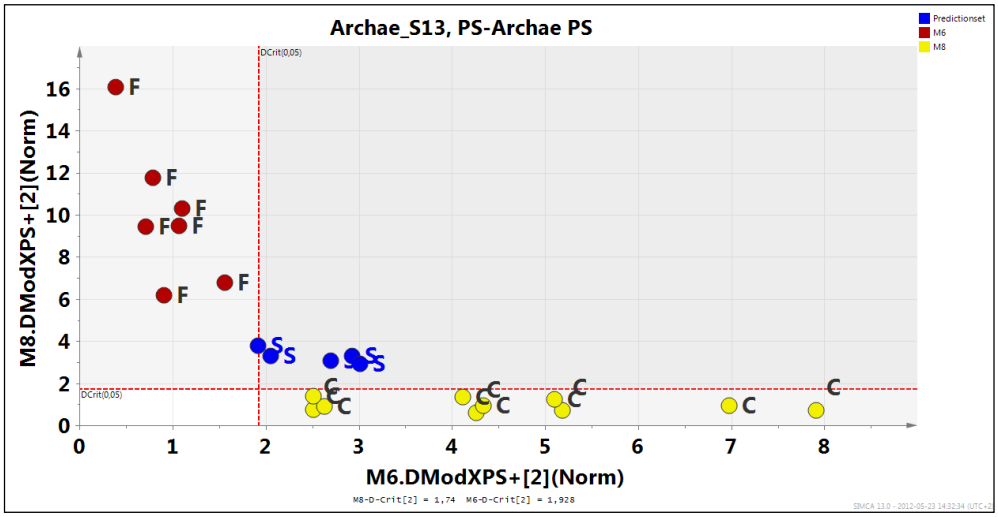


Figure 10.29: Coomans' plot with distance to the Control model plotted versus distance to the Feature model. This plot shows that two of the Site samples are close to fitting the Feature model.

10.6.3 PLS-DA of archaeology data

When PCA results in clustering of observations, it is sometimes worth trying to further resolve such groupings by means of PLS discriminant analysis. We created a dummy matrix of three Y-variables expressing class identity of the archaeological samples (cf. Figure 10.23). PLS-DA was then performed, which gave a four-component model with $R^2X = 0.97$, $R^2Y = 0.86$, and $Q^2Y = 0.69$.

In regular PLS analysis, score plots of the t/u type are of paramount interest. However, in PLS-DA, score plots of the t/t-type are also interesting, because one wants to overview the class discriminating ability of a developed model. Figure 10.30 shows the t_1/t_2 score plot of the PLS-DA model, and Figure 10.31 gives the corresponding weights ($w*c_1/w*c_2$).

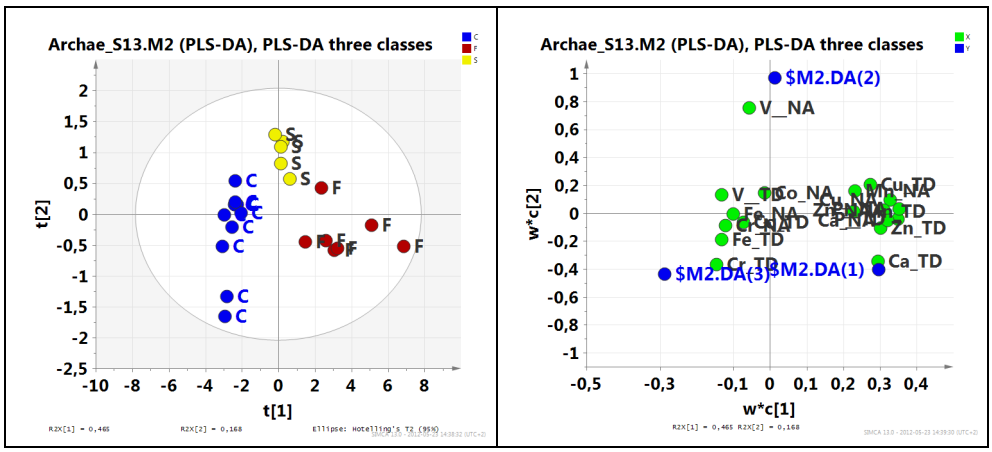


Figure 10.30: (left) PLS-DA t_1/t_2 score plot correlating to 72% of the Y-variance. The three classes of archaeological soil samples are clearly discriminated. Figure 10.31: (right) PLS weight plot ($w*c_1/w*c_2$) corresponding to Figure 10.30. Note the triangular distribution of the three dummy variables.

As shown by Figure 10.30, the three classes of archaeological soil samples are clearly separated. In order to understand which variables contribute to the separation, it is informative to consider the PLS weight plot displayed in Figure 10.31. Although the PLS model comprises four components, Figure 10.31 constitutes a sound basis for model interpretation as it represents 72% of the Y-variance.

The triangular distribution of the three dummy variables (denoted $DA1, $DA2, and $DA3) is expected because of their 0/1-nature. Chemical descriptors that are close to these dummy variables contribute strongly to the separation of the classes. Thus, in general, the Control observations have comparatively high levels of Cr, Fe, V, and Co, and low levels of Mn, Cu, Zn, Ca, and P. The Feature group of observations has a reversed pattern, and the Site samples fall between these two extremes.

More quantitative estimates of the discriminatory power of each chemical descriptor are obtained from the PLS regression coefficients or the VIP-parameter. The former are plotted if a single dummy variable is in focus, and the latter if an overall assessment of the variable importance is desired. Figure 10.32 shows the VIP plot of the PLS-DA model. Obviously, the vanadium content measured with the NA pre-treatment method displays the strongest discriminatory power.

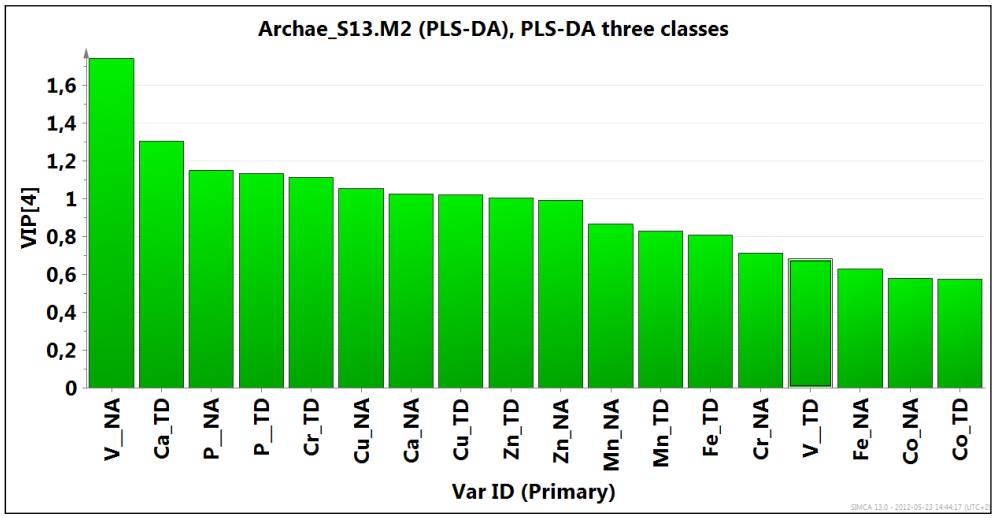


Figure 10.32: VIP plot of the PLS-DA model. The higher the numerical value, the stronger the discriminatory power of the chemical descriptor. The vanadium content, measured after NA pre-treatment, exhibits the strongest discriminatory power among the 18 descriptor variables.

10.6.4 Summary of ARCHAE application

The initial PCA model revealed strong groupings among the three categories of archaeological soil samples. It was found that the scores t_1 and t_2 jointly reflected the level of cultural disturbance in the soil samples. Furthermore, it was concluded that there was a larger difference between two main groups of elemental profiles than between the NA and TD pre-treatment techniques.

In the second stage, the SIMCA methodology was applied to the data. Thus, separate PCA models were fitted to the classes. These models were used to predict class membership of all 22 soil samples. It was concluded that the Site class was most homogeneous, in part because it contained only five members. The class model of the F-samples was found least specific.

In the last modeling stage, PLS-DA was attempted. A strongly significant PLS model was acquired, indicating that the 18 chemical variables indeed contained class separating information. The separation of the three classes was slightly superior compared with the previous modeling attempts. As a quantitative measure of the discriminatory power of the 18 chemical descriptors, the VIP parameter was used. It was inferred that the vanadium content, monitored after NA pre-treatment, was the descriptor carrying the most class discriminating information.

10.7 Questions for Chapter 10

1. What are the objectives of classification?
2. What are the most typical situations where classification may be applied?
3. How may classes be specified?
4. What is PARC? When is it useful?
5. What is the asymmetric case? How can it be conceptualized in process modeling?
6. Explain the SIMCA method. When is it useful?
7. What is the Coomans' plot?
8. What is PLS-DA? When is it useful?
9. Which kind of score plot is particularly relevant in PLS-DA?
10. Which PLS parameters may be interpreted in PLS-DA to reveal the discriminatory power of an X-variable?

10.8 Summary and discussion

10.8.1 How to formulate the classification problem

Classification and discriminant analysis is an important area of multivariate data analysis. Multivariate data analytical methods are able to utilize the information in highly multicollinear data for exploring within-class similarity and between-class diversity. This is of particular relevance in the areas of spectroscopy (e.g. NIR, UV, NMR) and “omics” data analysis (genomics, proteomics, metabonomics, etc.), where the number of variables can be very large and the correlation among them substantial. Classification analysis is often used in the pharmaceutical industry to verify that NIR spectra of products are similar to an appropriate reference spectrum stored in a reference data base, i.e., for verifying that the correct raw material was delivered from the supplier.

The largest obstacle in multivariate classification and discrimination is that of large inhomogeneities in classes believed to contain similar observations. Although multivariate methods may classify prediction set observations correctly despite inhomogeneities in the training data, the results are very difficult to interpret and utilize since the class concept is strongly linked to the idea of intra-class similarity. In SIMCA and PLS-DA, inhomogeneities and other anomalies in the training data are indicated by weak models and/or deviating observations in scores and model residuals.

We note that the specification of classes is part of the problem formulation. It must be done with caution and correspond to a real observable group similarity. The similarity within a class must be greater than the between-class similarity. Thus, an objective of the initial data analysis is to reveal whether a sound class division has been accomplished.

10.8.2 Main steps of SIMCA

In Chapter 10 we have outlined the SIMCA approach as a way to deal with classification when class membership is either not known at the beginning, or perhaps only vaguely specified. The most important steps of SIMCA can be summarized as follows:

(1) *Pre-processing of data*: With a least squares method like SIMCA, the results of the data analysis are sensitive to the scaling of the descriptor variables. As was pointed out in Section 3.3, scaling to unit variance and mean-centering should be done when there is no *a priori* knowledge about the importance of each variable. In both the IRIS and the ARCHAE applications, the data were mean-centered and scaled to unit variance. In addition, the X-data in the latter application were log-transformed to achieve more normally distributed data.

(2) *Overview of data*: The second step corresponds to an overview PCA or PLS model of the complete dataset. Such a model will indicate groups of observations, trends, outliers and other anomalies. This is the key step towards understanding how the observations should be divided into meaningful classes.

(3) *Disjoint class modeling*: In the next step, local models are fitted to the classes that were discernible in step 2. This makes it possible to better resolve heavily clustered data encountered in step 2, and to study in detail the finer structure in each class. It is important to use cross-validation to determine the appropriate complexity of each class model. If a class lacks structure, i.e., the data contain little information about the similarity of the samples in the class, this is seen as an estimated model complexity of zero components.

(4) *Model interpretation*: Each class model should be interpreted to understand which variables contribute to the modeling. Preferably, model interpretation should be consistent with existing fundamental chemical, biological, engineering, …, knowledge. Otherwise doubt is cast on the class division performed.

(5) *Classification of observations:* In this step the observations are projected onto the respective class model. The fit to each model is understood from the calculation of the distance to the model, DModX. Usually, an observation fits only one class, but occasionally conformity to more than one model is found. There are also observations that do not fit any class model. These observations should not just be ignored, but rather considered from the viewpoint of representing a new class of (challenging) properties.

(6) *Classification of new observations:* The ability of a class model to classify observations correctly ("classification rate") cannot be realistically assessed if only training set samples are considered. It is recommended that an independent prediction set of observations that have not influenced the model calculations be employed. This will enable a more realistic assessment of the classification performance of each model. In the IRIS dataset, an external prediction set was used. However, in the ARCHAE dataset this could not be done since only 22 observations were available.

10.8.3 When and when not to use PLS-DA

In addition to the SIMCA approach, the PLS-DA method was highlighted. Class memberships are explicitly given in PLS-DA by construction of an artificial Y-matrix of dummy variables. PLS-DA often rotates the latent variables, such that a maximum separation among classes is obtained.

A necessary condition for PLS-DA to work reliably is that each class is "tight" and occupies a small and separate volume in the X-space (see industrial example in Figure 10.33). As an illustration, Figure 10.33 shows a score plot where four types of wood fiber have been

separated using the combination of NIR-spectroscopy and PLS-DA. In PLS-DA, the number of modeled classes must not be too high. Experience shows that PLS-DA is useful with 2-4 classes, but when the number of classes exceeds four, discrimination results may become incomprehensible and difficult to overview. Moreover, when some of the classes are not homogeneous and spread significantly in X-space, the discriminant analysis does *not* work. Rather, other approaches, such as SIMCA must be used.

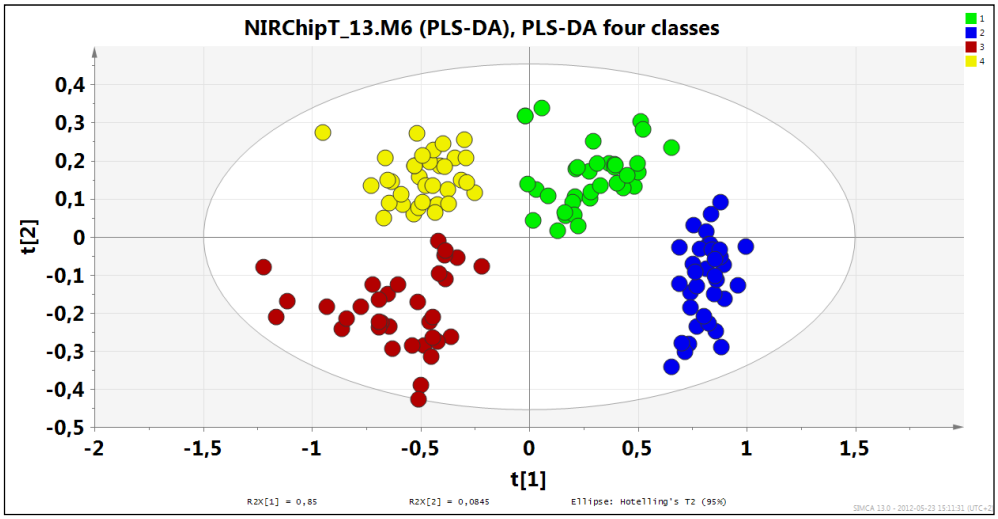


Figure 10.33: PLS-DA t_1/t_2 score plot showing the separation of four types of wood fiber based on information contained in their NIR profiles. The example stems from applications within the particleboard industry conducted by Casco Products, Sundsvall, Sweden.

10.8.4 Regular and combined DModX´s

In the classification and prediction phase, the DModX parameter is available in two versions, called regular (DModX) and combined (DModX+). The former parameter, DModX, measures the orthogonal distance of an observation to the model (line, plane, hyperplane, …). The latter parameter, DModX+, is based on the former but is augmented with a term measuring how far outside the model domain the classified or predicted observation falls. The relationship between the two statistics is summarized graphically in Figure 10.34.

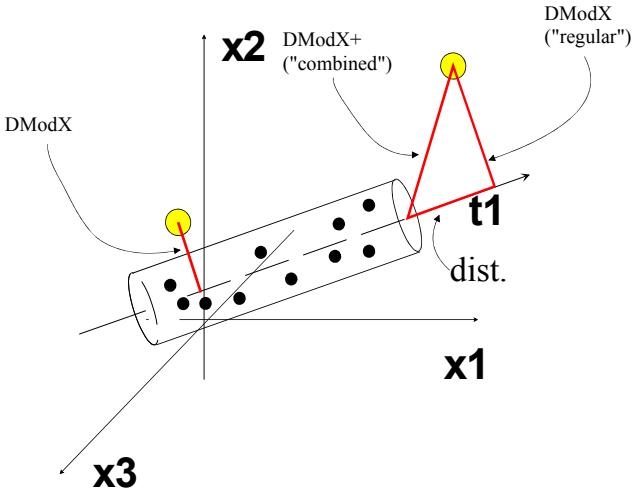


Figure 10.34: A graphical representation of the relationship between DModX and DModX+. DModX measures the orthogonal distance to the multivariate model, here represented as a one-dimensional model with the surrounding, cylindrical tolerance volume. A limitation of DModX is that it does not witness about whether a predicted observation falls inside or outside the range of the model. The extension of DModX, termed DModX+, provides this additional information. Mathematically, DModX+ is derived from Pythagora´s Theorem, according to: $DModX+ = [DModX^2 + dist^2]^{1/2}$.

The choice of which version of DModX to use depends on the objective and focus of the on-going investigation, i.e., (i) classification/prediction in a general sense, or (ii) contribution plotting in model residuals to reveal why an observation does not fit the multivariate model well. In the latter case, contribution plots in DModX will reveal which variables contribute to the large residual of a predicted observation. Here, the main interest lies in understanding what causes an observation to be far from (orthogonal to) the model (line, plane, hyperplane…), and it is not necessarily of relevance to know what features in the variables cause the observation to fall outside the model domain in the direction of the score vector(s). However, when classification is the sole objective, revealing whether an observation is located inside or outside the model tolerance volume is of paramount importance, and hence DModX+ is the appropriate parameter to use in this context.

Thus, in the SIMCA software, DModX is the default for contribution plots, whereas DModX+ is the basis for the Coomans' plot and the classification and misclassification lists (see further discussion in next section).

10.8.5 The relationship between the Coomans' plot and the Classification and Misclassification lists

In this section, the objective is to describe a few tools which are useful in routine application of prediction, classification and discrimination models. These tools are: the Coomans' plot, the classification list and the misclassification list. The basics of the Coomans' plot were introduced in Section 10.6.2 in connection with the ARCHAE example. Figure 10.35 shows the Coomans' plot based on two of the class-models developed for the IRIS example and using the training set as the prediction set. Figure 10.36 shows the corresponding plot for the real IRIS prediction set. The conclusion is that Iris Setosa is easy to separate, whereas Iris Virginica and Iris Versicolor overlap heavily.

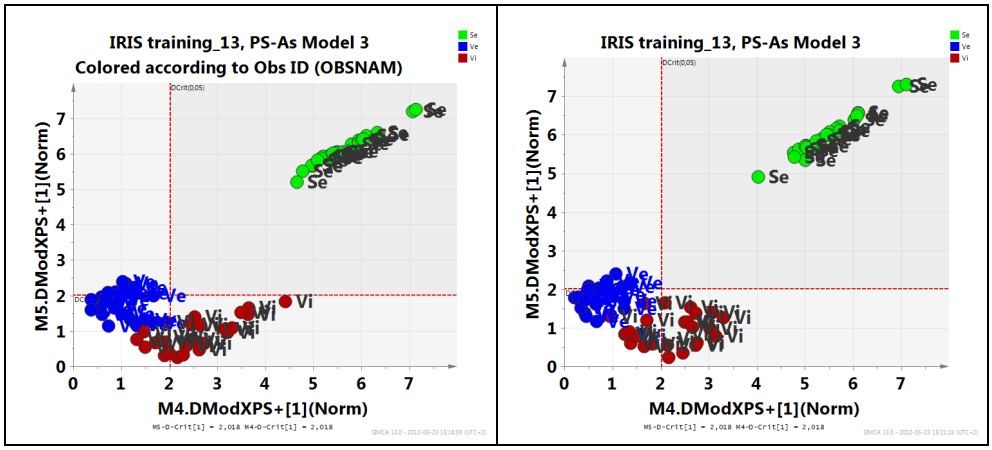


Figure 10.35: (left) Coomans' plot for the IRIS training set. The plot is based on the PCA models fitted to the Versicolor and Virginica classes. Figure 10.36: (right). Coomans' plot for the IRIS prediction set. The plot is based on the PCA models fitted to the Versicolor and Virginica classes.

The Coomans' plot is created by co-charting the DModX+ values of two models in the same scatter graph. The strength of the plot lies in its straightforward interpretation concerning the classification results of the two models in question. However, the fact that the plot only covers two classes at the same time also alludes to the limitation of this graphical tool.

Since having many classes is quite common, diagnostic tools capable of handling the classification results of multiple models are desirable. The classification list is one such tool. Basically, it is a table with the observations of the prediction set as rows, and the probability of belonging to the class models, PModX+, as columns.

Figure 10.37 shows the classification list using the IRIS training set as prediction set. Note, that the only reason for using the training set in this role is that this maneuver facilitates the explanation of the classification and misclassification lists. When the probability of belonging to a model is greater than 10%, the value is rendered against a green background.

When the probability is between 5 and 10%, the value is rendered against an orange background. When the probability is below 5%, the observation does not belong to that model and appears on a white background.

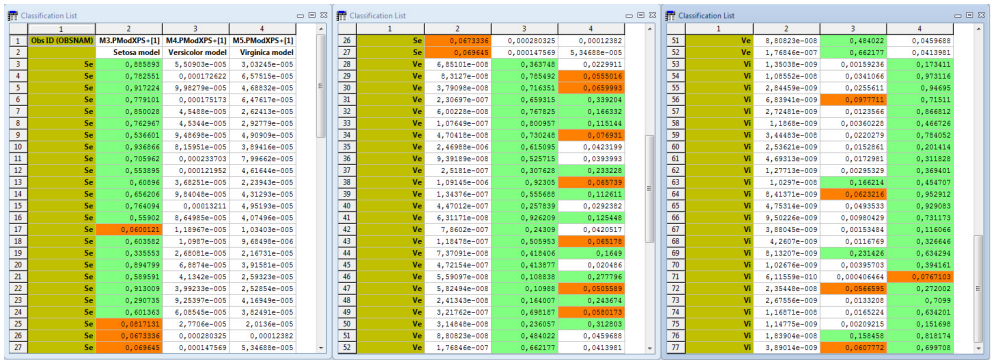


Figure 10.37: The classification list for the Iris training set. Each column provides the probability of class membership, PModX+, to the Setosa, Versicolor, and Virginica models. When the probability of belonging to a model is greater than 10%, the value is rendered against a green background. When the probability is between 5 and 10%, the value is rendered against an orange background. When the probability is below 5%, the observation does not belong to that model.

The misclassification list (Figure 10.38) is complementary to the classification list, but accounts for the classification results in a more compact format. Compared with the classification list, the misclassification list has tougher rules for class-assignment. When the probability of belonging to a model is greater than 10%, the observation is considered as a member of that particular class. When the probability is below 10%, the observation does not belong to that model.

Using the IRIS training set as prediction set, the misclassification list summarizes how well the three PCA class models classify the observations into the known classes.

The misclassification table is organized as follows (Figure 10.38):

- The first column is the number of observations known to belong to that class.
- Each row, starting at column 2, displays how the observations have been classified by the model.
- Hence all entries on the diagonal are observations that have been correctly classified.

In Figure 10.38, we can see that 25 out of 25 training set Setosas (Class 1) are correctly classified, 22 of 25 Versicolors (Class 2), and 25 of 25 Virginicas (Class 3). The overall classification rate is 96%

Misclassification Table for Model 3, 4, 5

	1	2	3	4	5	6	7
1		Members	Correct	1	2	3	No class (PModX+ < 0)
2	1	25	100%	25	0	0	0
3	2	25	88%	0	22	3	0
4	3	25	100%	0	0	25	0
5	No class	0		0	0	0	0
6	Total	75	96%	25	22	28	0

Figure 10.38: Misclassification list of the IRIS example, using the training set as the prediction set.

Now, that we have explained the basics of the classification and misclassification lists, we will apply the Setosa, Versicolor, and Virginica PCA class models to the 75 flowers residing in the IRIS prediction set. Recall that it is known that the prediction set contains 25 flowers of each species. The results obtained are summarized by the misclassification list seen in Figure 10.39. We can see that 25 observations are inside Class 1 (Setosa), 24 inside Class 2 (Versicolor) and 24 inside Class 3 (Virginica). The overall classification rate is 97.3%. A deeper discussion of these results was given in Section 10.4.6.

Misclassification Table for Model 3, 4, 5

	1	2	3	4	5	6	7
1		Members	Correct	1	2	3	No class (PModX+ < 0)
2	1	25	100%	25	0	0	0
3	2	25	96%	0	24	1	0
4	3	25	96%	0	1	24	0
5	No class	0		0	0	0	0
6	Total	75	97,33%	25	25	25	0

Figure 10.39: Misclassification table of the IRIS example using the true prediction set.

10.8.6 Related methods

Of course, there are numerous other methods amenable to classification. The oldest, known as linear discriminant analysis (LDA) and linear learning machine (LLM), work well provided that all classes are strictly homogeneous and that the number of X-variables is much smaller than the number of samples in the training set [Albano, *et al.*, 1984]. Since both conditions are violated with multivariate and multicollinear data, other methods are usually preferred.

Naes and Indahl worked out a unified description of classical classification methods for multicollinear data [Naes and Indahl, 1998]. They showed that a number of the established methods for classification can be embraced by the same basic mathematical framework. Some of the methods addressed are linear discriminant analysis, quadratic discriminant analysis, and their modifications to comply with multicollinearity, DASCO, and SIMCA. The reader is encouraged to consult the cited paper for further reading. The book by Massart *et al.* [Massart, *et al.*, 1988] also contains discussions on the subject of classification and discrimination.

11 Identification of discriminating variables

11.1 Objective

In this chapter we introduce orthogonal partial least squares discriminant analysis (OPLS-DA). OPLS-DA has the same interpretational advantage over PLS-DA as OPLS over PLS, namely that it can better cope with a high degree of orthogonal variation in the X-block. OPLS-DA is especially suited for highlighting class discriminating variables in the two-class problem. With OPLS-DA comes a number of tailor-made plots assisting in the identification of such discriminating variables. These plots, denoted the S-plot, S-line plot, and SUS-plot, are exemplified.

11.2 Introduction

In Chapter 10 we introduced the methods SIMCA and PLS-DA and documented their use in classification and discrimination modeling. As afore-mentioned, SIMCA is a PCA based method, which consists of computing a local PCA model for each class. It should be noted that when a single PCA is used in a PARC problem with a set of observations representing one or several classes, knowledge related to class membership is *not* used to find the location of the principal components. PLS-DA, on the other hand, explicitly takes into account the class membership of observations through the artificial Y-matrix consisting of 1/0 dummy variables. This makes it possible to accomplish a rotation of the projection to give latent variables better focus on class separation.

However, the applicability of PLS-DA is critically influenced by the amount of orthogonal variability in the X-block. As was shown in Chapter 5, the interpretability of a PLS model is negatively affected when orthogonal variability is the dominating structure in X. In such a case it is not fully clear what the PLS loadings mean, since they reflect combinations of predictive and orthogonal variabilities. Since the estimation algorithm is the same in PLS-DA, this method suffers from the same limitations as PLS when the amount of orthogonal variability is high. This means there is a risk that it can be difficult to reliably identify class discriminating variables using PLS-DA.

As we shall see in this chapter, the identification of class discriminating variables is facilitated when replacing PLS-DA by OPLS-DA, particularly for the two-class discriminant problem. The reason for this is that in a two-class discriminant problem, there can, by theory, only be one single predictive component in the OPLS-DA model. Hence, to identify the class discriminating variables, the analyst only has to construe one single predictive loading vector.

The OPLS-DA method and its applicability in highlighting discriminating variables between classes will be illustrated using three examples. The first example is a tutorial example that is selected because it is easy to understand its meaning. The second example comes from plant metabolomics and the third is a process example. These examples will be used to introduce three tailor-made plots called S-plot, S-line plot and SUS-plot, which are dedicated to uncovering the most reliable class discriminating variables. Moreover, the S- and SUS-plots can be used to display the orthogonal loadings to learn more about the orthogonal information.

11.3 A tutorial example: Distilled Liquor

11.3.1 Introduction

Complex liquid samples can be characterized, compared and classified with the help of a non-selective analytical method, for instance one which takes advantage of the samples' ability to absorb visible light. From the characterization of samples of known origin, predictive models can be built and tested with new samples of unknown composition. In this tutorial, a range of distilled spirits are investigated using vis-spectroscopy. We are grateful to professor Johan Trygg and his colleagues at Umeå University for granting us access to this dataset.

The objective of this exercise is to provide an illustration of multivariate characterization based on spectral data. To this end, spectra measured on a set of alcoholic spirits will be used. The spirits can be compared and classified by investigating if there are clusters relating to, for example, product type or country of origin. A growing problem in the beverage and brewing industry is fraud and plagiarism; see for example in the reference [Jos, *et al.*, 2004] in which sparkling wines (champagne and cava) were differentiated using a multivariate model based on mineral content profiles. Chemometric methods can greatly assist in identifying incorrectly labeled or fake products.

For each sample (spirit), the visible spectrum (200–600 nm) was acquired using a Shimadzu spectrometer. Signal amplitude readings were taken at 0.5 nm intervals yielding 801 variables. To illustrate the OPLS-DA method we have selected a subset comprising 18 samples distributed as 10 blended whiskies and 8 single malts, all from Scotland. The raw data of these 18 samples are plotted below (Figure 11.1).

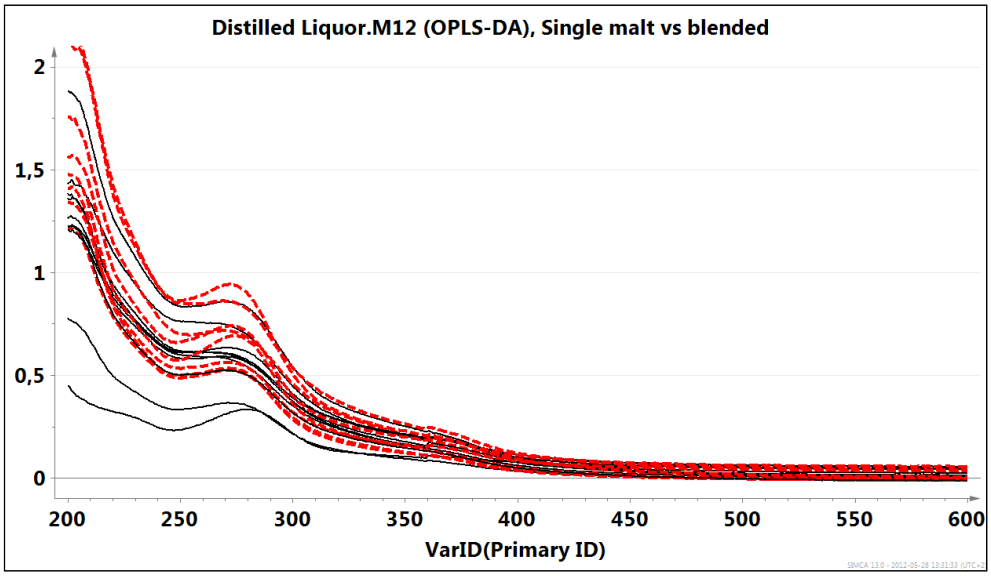
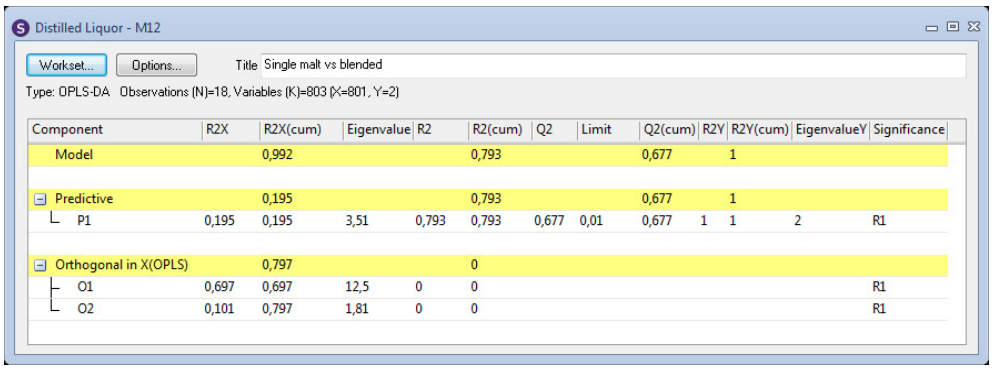


Figure 11.1: Spectral data for the 18 whisky samples. Red dashed lines represent blended whiskies (10 varieties) and solid black lines single malts (8 varieties). The spectral window is between 200 and 600 nm and corresponds to 801 variables.

11.3.2 Two-class discriminant modeling

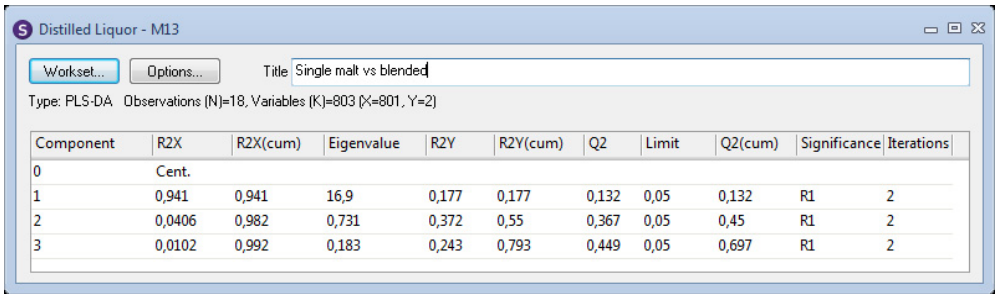
Prior to the multivariate modeling, the spectral data were centered but not scaled. Note that different approaches to scaling and centering of multivariate data are discussed in more detail in Chapter 13. OPLS-DA was used to discriminate between the single malts and the blended whiskies. As shown by Figure 11.2, the obtained model consisted of 1 predictive and 2 orthogonal components. The model finds 19.5% of the spectral variance to be predictive to Y (the matrix of artificial 1/0 dummy variables) and 79.7% of the spectral variance to be orthogonal to Y. An analogous PLS-DA model fitted to the same data

comprised 3 components with R2X = 99.2% and R2Y = 79.3%, i.e., as expected, identical explained variabilities to the OPLS-DA model (Figure 11.3).



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,992			0,793			0,677		1		
Predictive		0,195			0,793			0,677		1		
P1	0,195	0,195	3,51	0,793	0,793	0,677	0,01	0,677	1	1	2	R1
Orthogonal in X(OPLS)		0,797			0							
O1	0,697	0,697	12,5	0	0							R1
O2	0,101	0,797	1,81	0	0							R1

Figure 11.2: Model window of the 1+2 OPLS-DA model discriminating between single malts and blended whiskies. The model explains 99.2% of the systematic variability in the spectral data and 79.3% of the variability in the artificial Y-data. Only 19.5% of the spectral variability is predictive to Y.



Component	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.									
1	0,941	0,941	16,9	0,177	0,177	0,132	0,05	0,132	R1	2
2	0,0406	0,982	0,731	0,372	0,55	0,367	0,05	0,45	R1	2
3	0,0102	0,992	0,183	0,243	0,793	0,449	0,05	0,697	R1	2

Figure 11.3: Model window of the 3 component PLS-DA model fitted to the two-class problem of separating single malts and blended whiskies from one another. This model has identical overall statistics to the OPLS-DA model.

A comparison of the two first score vectors of the OPLS-DA and PLS-DA models (Figures 11.4 and 11.5) document the impact of OPLS-DA in a two-class discriminant problem; the predictive information of Y in X is concentrated in the first predictive component. The variation in X which is unrelated to Y is put in the second and orthogonal component. In other words, in Figure 11.4 the horizontal direction corresponds to between class variability and the vertical direction to within class variability. This partitioning of the X-data results in improved model transparency and interpretability, since it is better understood what the different components of the OPLS-DA model mean. PLS-DA, on the other hand, confuses the predictive and orthogonal spectral variabilities.

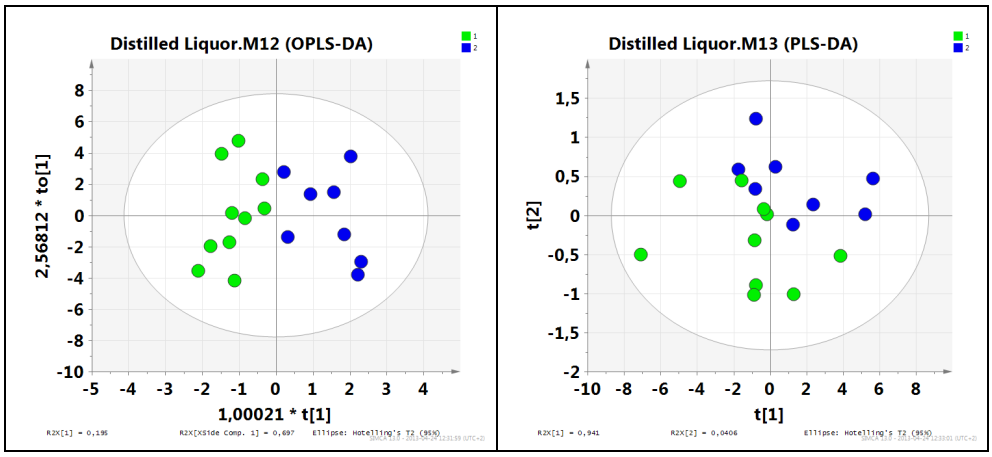


Figure 11.4: (left) OPLS-DA score plot with the first orthogonal score vector plotted against the single predictive score vector. Green points denote blended whiskies and blue dots denote single malts. The single predictive component separates between the two classes. Figure 11.5: (right) PLS-DA score plot with the second component plotted versus the first. Coloring as in Figure 11.4.

In order to uncover which parts of the spectral window carries the class separating information, it is enough to plot the loading vector of the single predictive component of the OPLS-DA model. This is done in Figure 11.6. The plot in Figure 11.6 is called the S-line plot in the SIMCA software. The plot displays the predictive loading in a form resembling the original spectra and is colored according to the correlation scaled loading. More information about this plot, and others geared for the same objective, is given below in Section 11.4.3.

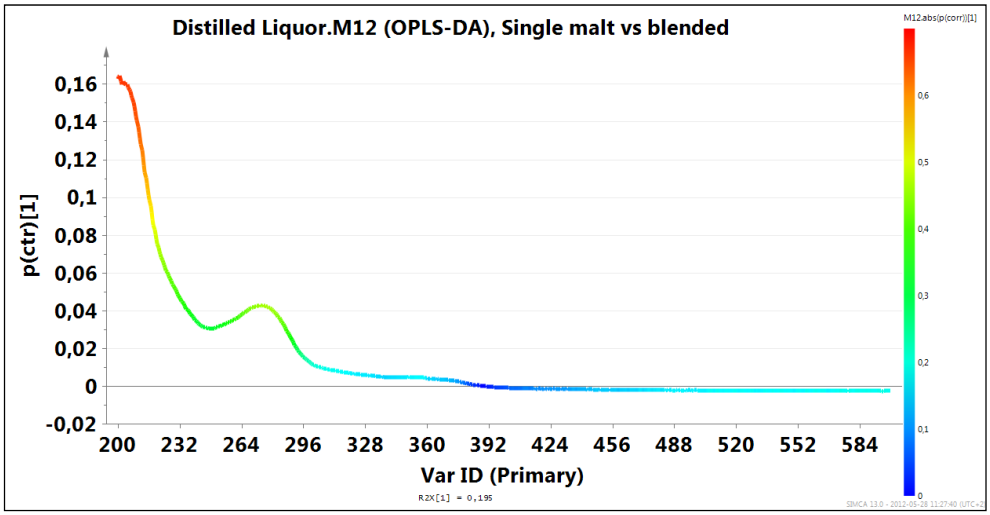


Figure 11.6: S-line plot of the single predictive loading of the OPLS-DA model. High numerical loading values combined with a red color encode spectral regions of strong class discriminatory power between the single malts and the blended whiskies.

Beyond interpreting the predictive variability in the measured spectra one should of course not forget to interpret the orthogonal variability to enhance the total understanding of the studied problem. Such interpretations are considered for the third example (Section 11.5).

11.4 Omics example: Poplar

11.4.1 Introduction

The second example concerns a study of genetically modified poplar plants by GC/MS metabolomics. We will use it to illustrate how OPLS-DA may be used in a multi-class problem in which there is an apparent reference class (control class). This will lead to a discussion on the three dedicated plots mentioned in Section 11.2, i.e., the S-plot, the S-line plot and the SUS-plot.

The experimental objective was to investigate two genetic modifications, i.e., up-regulation and down-regulation within the PttPME1 gene. This gene is involved in the production of pectinmethylesterase, PME, which is an enzyme that de-esterifies methylated groups within the pectin. Pectin is complex polymer and will not be analyzed by this type of technique. Nevertheless, the metabolic profile was of interest as both lines indicated several symptoms of oxidative stress response.

In total, the data set contains N = 26 observations (plants) and K = 80 variables (resolved and integrated GC/MS profiles). The observations (plants) are divided in three groups (“classes”):

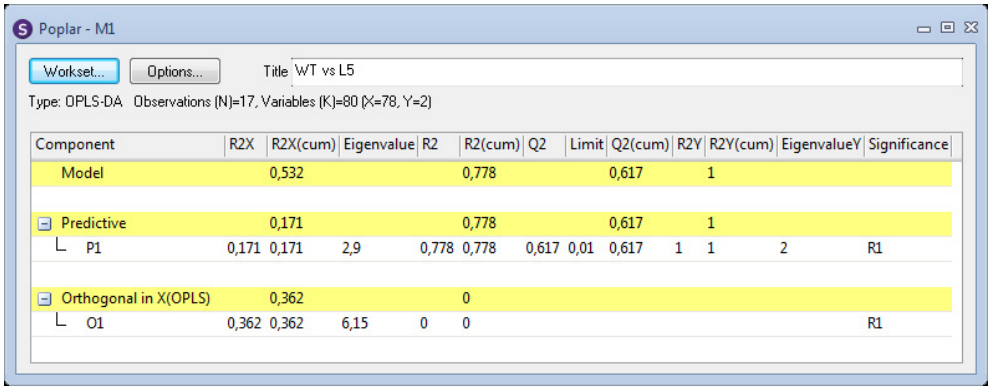
- Wild type plant, 10 plants “WT”
- PttPME1 down-regulated poplar, 7 plants “L5”
- PttPME1 up-regulated poplar, 9 plants “2B”

The GC/MS data are three-way by nature with time, absorbance and mass dimensions. In this example the three-way data have been pre-processed by Hierarchical Multivariate

Curve Resolution, HMCR [Jonsson, *et al.*, 2006]. HMCR resolves the chromatographic profiles, calculates the area of the resolved profiles and generates the corresponding mass spectrum. The resolved mass spectrum can be subjected to a library search and the compound can thus be identified. A more elaborate account of the experimental conditions is found in the original reference [Wiklund, *et al.*, 2008].

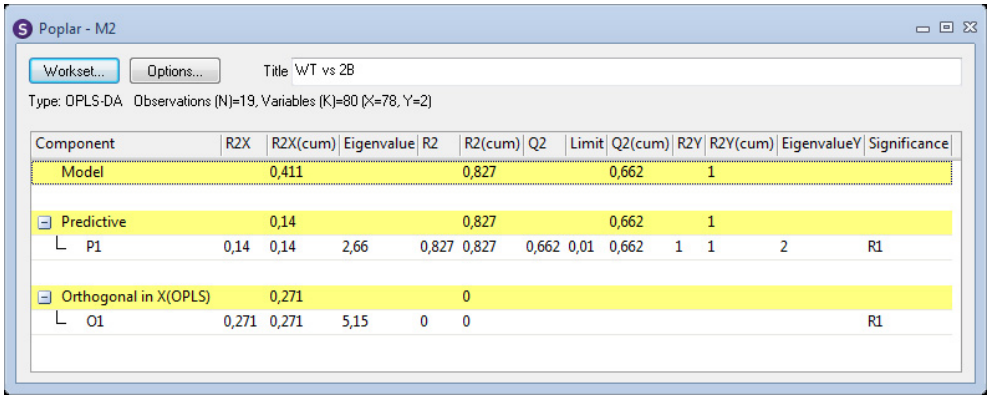
11.4.2 OPLS-DA models between control and treatment groups

In order to compare the control and treatments groups, two OPLS-DA models were computed. Prior to the modeling, the mass spectral data were scaled using Pareto scaling. Pareto scaling is discussed in Chapter 13. The first model consisted of comparing the wildtype with the down-regulated poplar samples (WT vs. L5) and the second model consisted of comparing the wildtype with the up-regulated poplar samples (WT vs. 2B). Figures 11.7 and 11.8 show the Model windows of the respective model.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,532			0,778			0,617		1		
Predictive		0,171			0,778			0,617		1		
P1	0,171	0,171	2,9	0,778	0,778	0,617	0,01	0,617	1	1	2	R1
Orthogonal in X(OPLS)		0,362			0							
O1	0,362	0,362	6,15	0	0							R1

Figure 11.7: Model window of the 1+1 OPLS-DA model comparing wildtype with down-regulated poplar samples.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,411			0,827			0,662		1		
Predictive		0,14			0,827			0,662		1		
P1	0,14	0,14	2,66	0,827	0,827	0,662	0,01	0,662	1	1	2	R1
Orthogonal in X(OPLS)		0,271			0							
O1	0,271	0,271	5,15	0	0							R1

Figure 11.8: Model window of the 1+1 OPLS-DA model comparing wildtype with up-regulated poplar samples.

As shown by Figures 11.7 and 11.8 both models contain one predictive and one orthogonal component. We can see that 17.1% and 14.0% of the X-variabilities are predictive to Y, i.e., correlated to class separation. Even higher fractions of the systematic X-variabilities are orthogonal (uncorrelated) to class separation. These numbers amount to 36.2% and 27.1%. The high R2(cum) for both models suggest good class separation. That this is indeed the case is evidenced by the score plots seen in Figures 11.9 and 11.10.

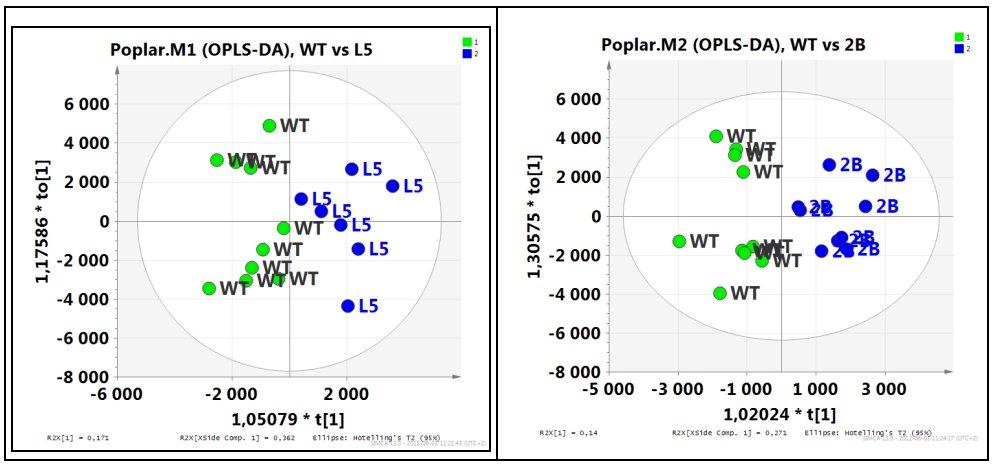


Figure 11.9: (left) OPLS-DA score plot of the model discriminating between wildtype (WT) and down-regulated (L5) poplar samples. The horizontal axis corresponds to between class variability and the vertical axis to within class variability. Figure 11.10: (right) OPLS-DA score plot of the model discriminating between wildtype (WT) and up-regulated (2B) poplar samples. The horizontal axis corresponds to between class variability and the vertical axis to within class variability.

Because the class separation is confined to the predictive component, it is sufficient to interpret the corresponding predictive loading vector to identify the variables with the best discriminatory power. The SIMCA software has three dedicated plots, called the S-plots, which provide visualization of the OPLS/OPLS-DA predictive component loading to facilitate model interpretation. The plots are only available for OPLS models with one y-variable and OPLS-DA models with two classes. These plots are described in the next section.

11.4.3 The S-plots

The first plot is the S-plot®, which is used to visualize both the covariance and the correlation structure between the X-variables and the predictive score t_1. Thus, the S-plot is a scatter plot of the p_1(corr) vs. p_1 vectors of the predictive component [Wiklund, *et al.*, 2008; Fotakis, *et al.*, 2013]. The former vector is the normal loading re-expressed as the correlation (running between -1 and +1) between each X-variable and the score vector t. With Pareto scaling, or with centering but no scaling, this plot often takes the shape of the letter 'S'. Hence, its name. X-variables situated far out on the wings of the S combine high model influence with high reliability and are of relevance in the search for discriminating variables or biomarkers that are up- or downregulated. In other words, the S-plot visualizes both magnitude (intensity) and reliability. In spectroscopic data the peak magnitude is important as peaks with low magnitude are close to the noise level and thus have a higher risk for spurious correlation. High reliability implies an X-variable with a strong correlation with t_1 and improved certainty as a discriminating variable or putative biomarker. It should be noted that the principle of the S-plot applies to the orthogonal information, as well. When the orthogonal information is a significant part of the model, it is recommended to examine it using the S-plot.

The second plot among the S-plots is the S-line plot. The S-line plot is tailor-made for NMR spectroscopy data, but it may be used for other types of data, as well. It visualizes the p_1(ctr) loading colored according to the absolute value of the correlation scaled loading, p_1(corr). The S-line plot thus displays the predictive loading in a form resembling the original spectra, colored according to p(corr). Thus, the S-line plot is conceptually similar to the STOCSY® plot pioneered by the Nicholson group at Imperial College, London, UK [Cloarec, *et al.*, 2005]. However, it should be noted that the coloring principle is slightly different between the two plots.

The last plot in the group of S-plots is called the SUS-plot, where SUS is the acronym for Shared and Unique Structures. The SUS-plot is a scatter plot of the p_1(corr) vector from two separate OPLS or OPLS-DA models [Wiklund, *et al.*, 2008]. If two OPLS or OPLS-DA

models have similar profiles (they capture similar relationships between the X-variables and the Y-variables) the X-variables will line up along the diagonal running from the lower left corner to the upper right corner. Such variables represent the Shared structure among the two compared OPLS or OPLS-DA models. Conversely, X-variables that are not located along this diagonal, e.g. in the mid left part, represent structures that are Unique to either of the two models compared.

11.4.4 Interpretation using the S-plots

In this section, we shall work out the interpretation of one of the models using the various S-plots. It should be noted that, given the nature of the data (mass spectra), the use of the S-line plot is a bit esoteric, and usually one would not use it in connection with mass spectra. The S-line plot is especially useful with NMR-data. However, since our objective is of a tutorial nature and we desire to exemplify the three S-plots in a coherent manner, we will present model results using all three S-plots.

The model we zoom onto is the first model, the one comparing wild-type with the down-regulated poplar samples. Figure 11.11 provides the S-plot and three variables (sucrose, malic acid and phosphoric acid) have been marked. Malic acid and phosphoric acid are positioned on the right-hand side of the 'S'. Hence, they have strong model contribution and high statistical reliability as discriminating variables (putative biomarkers in this case).

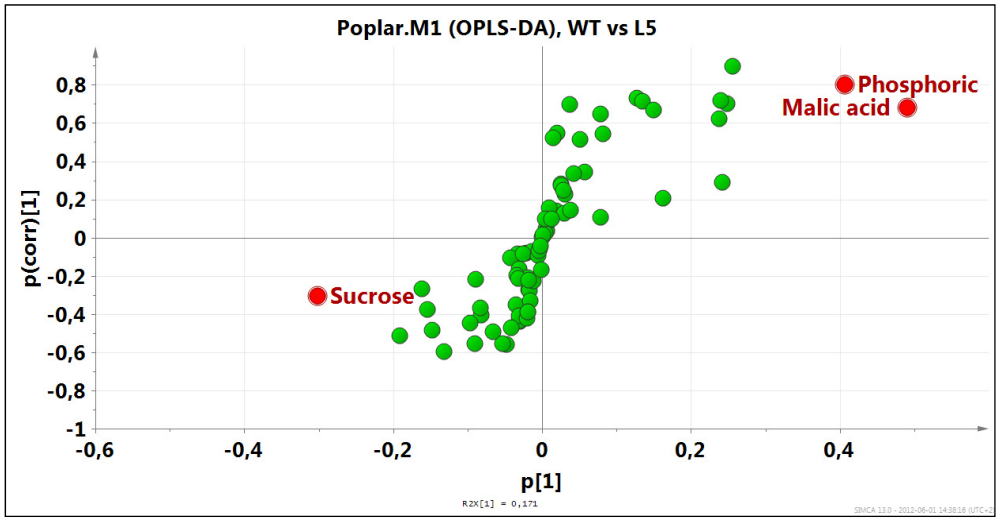


Figure 11.11: S-plot of the OPLS-DA model between WT and L5.

At the left-hand side of the 'S' we find sucrose, which has a strong contribution to the model, but at the same time lower statistical reliability than malic acid and phosphoric acid. If coupled to a column plot of p_1, the information from the S-plot is more readily understood. Figure 11.12 shows this column plot, in which confidence intervals, as estimated by jack-knifing, are superimposed on top of the loadings. Evidently, sucrose is not a reliable discriminating variable, whereas malic acid and phosphoric acid are. Always remember there can be a difference between statistical significance and real importance from a practical perspective.

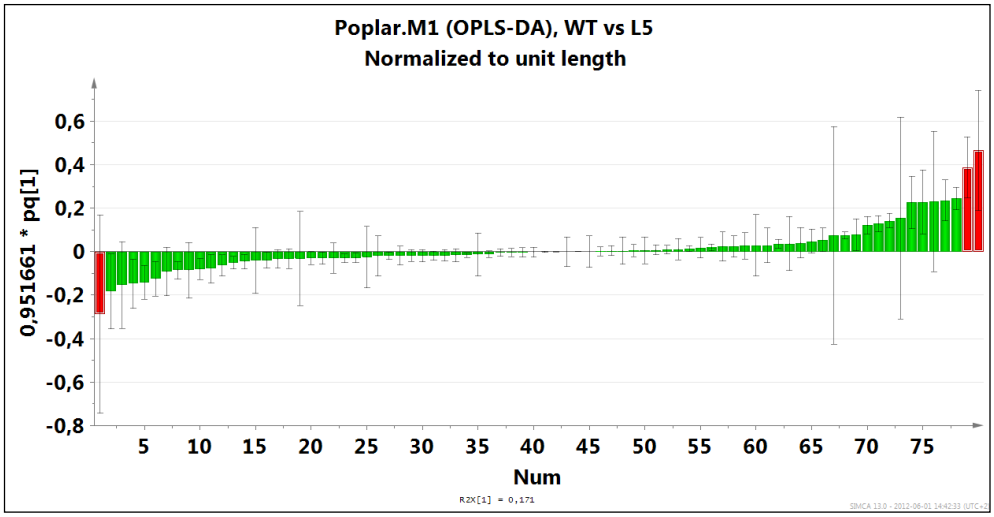


Figure 11.12: Column plot of the predictive loading of the OPLS-DA model between WT and L5. The same three variables have been marked as in Figure 11.11. Sucrose is the leftmost column and phosphoric and malic acids are the two rightmost ones. The uncertainty in each loading value is indicated by approximate 95% confidence intervals estimated via jack-knifing.

The S-plot is a practical and reliable way to identify important discriminating variables in the two-class discriminant problem. One limitation of the S-plot, however, is that it does not consider any physical ordering structure that might exist among the variables. As discussed in the foregoing section, the S-line plot is a plot that indeed does take advantage of e.g. the spectral order that prevails in many types of data. Figure 11.13 shows what the S-line plot looks like in the case we are currently investigating. The same three variables have been marked as in Figure 11.11 and 11.12. However, as emphasized above, one would typically not produce such a plot when working with mass spectral data; rather, the plot targets NMR data and other spectral data. In the S-line plot the relevance to the model is indicated by the signal amplitude and the significance by the color. Strongly discriminating variables therefore combine high numerical loading value (in the absolute sense) and red to orange-red color.

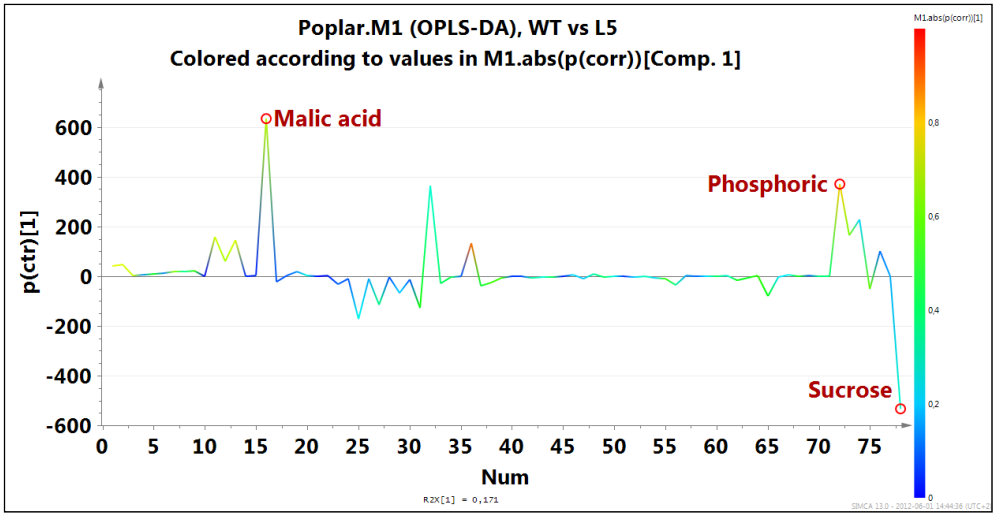


Figure 11.13: S-line plot for the OPLS-DA model between WT and L5. Sucrose, phosphoric acid and malic acid have been marked for the purpose of comparing with Figures 11.11 and 11.12.

Figure 11.14 shows the SUS-plot for the two OPLS-DA models. Seven variables (metabolites) have been marked for reference purposes. The SUS-plot shows that many variables (metabolites) have the same effect on both L5 and 2B. All the variables (metabolites) on the diagonal running from the lower left-hand to the upper right-hand corner have same effect in both plants, e.g. linoleic acid (down-regulated) and phosphoric

acid (up-regulated). The farther out from the plot origin, the stronger the discriminant effect of that variable between the treatment and control groups.

Conversely, any metabolite on the opposite diagonal would correspond to metabolites that have the same effect but in opposite directions in both transgenic plants. In this case, however, there are no such metabolites. Furthermore, discriminating variables (metabolites) that have an effect in one model, but not in the other, are located close to the X- and Y-axes and far out towards their ends. In this case, Glucaric acid and Quinoc acid are discriminating variables in the first model (WT vs. L5) and Inositol and beta-Sitosterol in the second model (WT vs. 2B). Finally, we have marked Glutamine which has very little discriminating power in both models; this is inferred from its proximity to the plot origin.

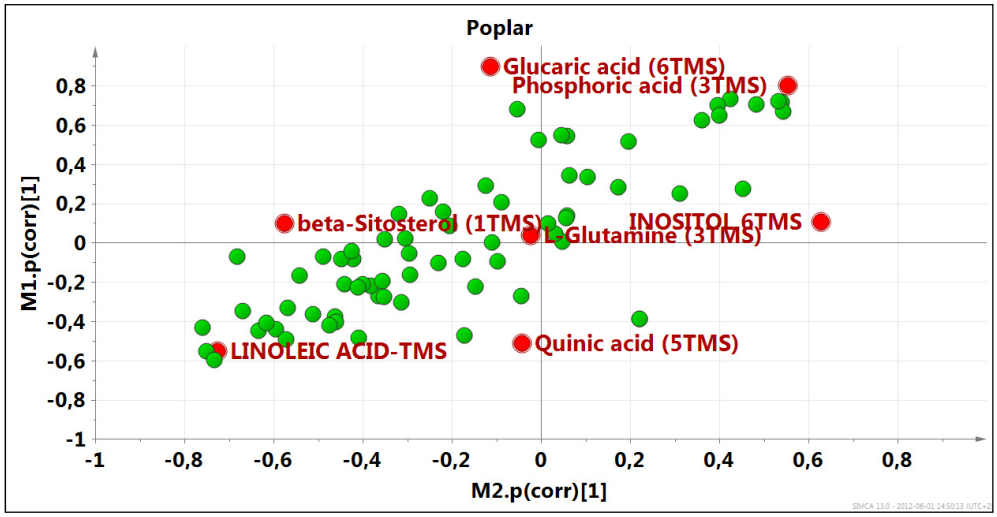


Figure 11.14: SUS-plot created from the predictive component of the two OPLS-DA models. Seven variables have been marked for discussion.

11.4.5 Summary of the Poplar example

The Poplar example delineates how to analyze a multi-class problem in which one of the classes has a special status as a reference group, standard group or control group. It is then very straightforward to set up several two-class discriminant models, in which the reference group is always contrasted against the modified group, or the group containing observations subject to some kind of "treatment". In such a case, the various S-plots can be constructed to highlight the most discriminating variables.

We emphasize that the S-plots may be created using PCA or PLS-DA models, however, only if a clear class separation is seen in the first component in the corresponding score scatter plot. Unless there is a clear discrimination among the groups of observations along the first component, the vectors p_1 and $p_1(\text{corr})$ will be confounded by variation that is unrelated to class separation, a fact that eventually will lead to a misleading interpretation.

When the S-plots are employed properly, we obtain a list of discriminating variables or, in metabolomics, potential biomarkers, which are statistically significant and which separate one class from another. It should be noted that in metabolomics such biomarkers are statistically significant, but not necessarily biochemically significant. They may have biochemical significance and they may be the biomarkers we are interested in, however, this must be established through extensive complementary experimental testing. Metabolomics allows the user to mine complex data and provides information which allows us to propose intelligent hypotheses. OPLS-DA is an excellent tool for this purpose.

11.5 Main example: Wafers

11.5.1 Introduction

The third example is the main example and here we shall also pay attention to the orthogonal variation. It is a process example stemming from the production of a computer chip. Making such a chip starts with the building of layers, both metal and insulator, onto a silicon wafer. Each wafer goes through hundreds of process equipment before it can be cut down into the computer chips that we use in our cell phones, laptops, and game consoles.

We here analyze data coming from an oxide chemical vapour deposition (CVD) process. The dataset contains three months of process data collected across three similar process chambers. The three process chambers are supposedly identical in their performance and it is of relevance to investigate whether this is the case.

The CVD is a multistage process and can hence be treated like a batch process consisting of many phases. In the current case, the deposition process consisted of 12 phases or steps. For each step, trajectories of many process variables were monitored, including gas flows, temperatures, pressures and equipment settings like angles and positions. Since we have not yet familiarized ourselves with how multivariate tools like PCA, PLS and OPLS can be adapted to suit the needs of evolving batches and their variable trajectories, we shall here analyze average values, formed on a phase-by-phase basis, of the process variables. The multivariate approach to batch statistical process control is discussed in Chapter 16.

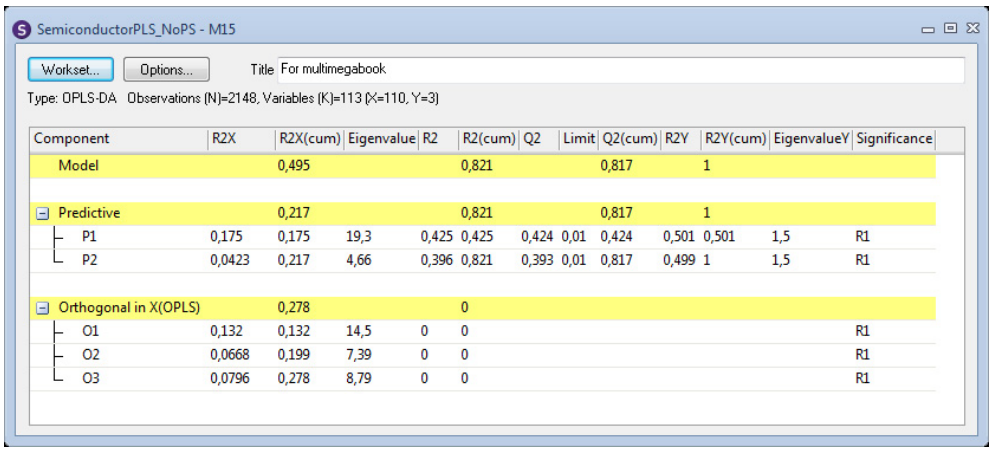
The Wafer dataset contains 2148 observations (wafers) processed by the three chambers. The distribution of the wafers is as follows: 675 wafers were processed by chamber A, 768 by chamber B, and 705 by chamber C. Thus, the distribution of wafers among the three chambers is well-balanced and this is a prerequisite for obtaining reliable class separation results. In order to describe the manufacturing of the wafers, a total of 110 process variables were monitored. In some phases (steps), seven process variables were measured and in some phases as many as 12 process variables were registered.

11.5.2 OPLS-DA model between the three chambers

In order to encode class (chamber) information, an artificial Y-matrix consisting of three 1/0 dummy variables was created. Recall that the principles behind such a Y-matrix were shown in Figure 10.23. Since this Y-matrix has rank 2, the OPLS-DA model will have two predictive components. Figure 11.15 shows the Model window of the resulting 2+3 OPLS-DA model. The model accounts for 49.5% of the X-variation and 82.1% of the Y-variation, which is a very good result considering that it covers 2148 observations (wafers).

Apparently, only about one-fifth (21.7%) of the variation in the process variables is correlated to systematic differences between the chambers. This can be seen as an encouraging result, which indicates that the major part of the variation in the process factors indeed correlate to other features, such as various product qualities. The OPLS-DA algorithm estimates that another 27.8% of the variation in the process variables is systematic and is thus descriptive for within chamber variability.

Moreover, if the late PCA step of the O2PLS technique is executed (see discussion in Section 6.3.2) -- to convert the OPLS-DA model to the corresponding O2PLS-DA model -- four PCA-based orthogonal-in-X components would account for an additional 18.2% of the X-variation. This means that 67.8% of the variation in the process variables is systematic. For reasons of brevity, however, we here interpret only the 2+3 OPLS-DA model, the numerical details of which are shown in Figure 11.15.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,495			0,821			0,817		1		
Predictive		0,217			0,821			0,817		1		
P1	0,175	0,175	19,3	0,425	0,425	0,424	0,01	0,424	0,501	0,501	1,5	R1
P2	0,0423	0,217	4,66	0,396	0,821	0,393	0,01	0,817	0,499	1	1,5	R1
Orthogonal in X(OPLS)		0,278			0							
O1	0,132	0,132	14,5	0	0							R1
O2	0,0668	0,199	7,39	0	0							R1
O3	0,0796	0,278	8,79	0	0							R1

Figure 11.15: Model window of the OPLS-DA model separating the three chambers. It can be seen that 21.7% of the variability among the process variables correlate to the three artificial Y-variables encoding the class information. The model explains 82.1% of the class membership information. Additionally, the model contains three orthogonal-in-X components that capture 27.8% of the variability among the process variables that is uncorrelated to the Y (the class membership information).

11.5.3 Interpretation of the predictive components

The scores and loadings of the two predictive components are plotted in Figures 11.16 – 11.18. The scores in Figure 11.16 show a clear discrimination among the three chambers. This does not come as a surprise considering the high R2 of 82.1%. Roughly 1770 of the 2148 wafers are positioned in the three main clusters that encapsulate the plot origin. This corresponds to approximately 82.4% of the total number of wafers; thus, a value very well matching the R2 of the model. This means that roughly 380 wafers are distributed near the origin of the score plot and for these there are no differences in process variables due to choice of process chamber.

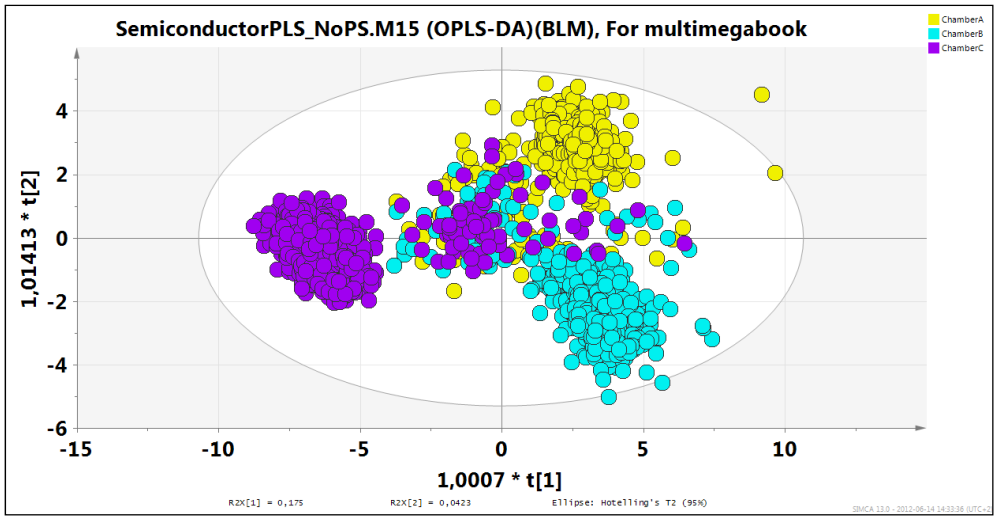


Figure 11.16: Score plot of the two predictive components. Each point is one observation (wafer). Coloring is according to which chamber was used in the manufacturing of each wafer.

Figure 11.17 provides the same score plot as Figure 11.16, the difference being that the coloring principle is according to product quality. Product quality is in this case oxide layer thickness. Evidently, wafers processed through chamber B have generally slightly thinner oxide layers than the wafers processed by chambers A and C. It is important to note that information about product quality (oxide layer thickness) was with-held from the modeling step and so it has not influenced the distribution of the observations (wafers) is the score space.

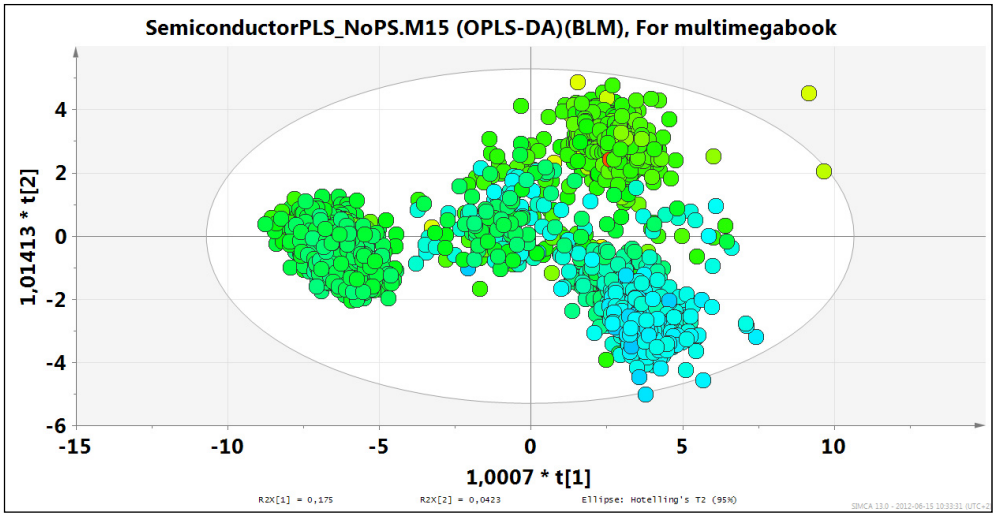


Figure 11.17: Plot analogous to Figure 11.16 but colored according to product quality (oxide layer thickness). Wafers processed by chamber B generally have slightly thinner oxide layers than wafers produced by chambers A and C.

Although there are indications that wafers processed by chamber B have slightly thinner oxide layer than others, the performance of chamber C is most dissimilar (to chambers A and B). This inferred from the fact that the first predictive component accounts for 17.5% of the variation in the process factors, whereas the corresponding number for the second predictive component is merely 4.2%. Hence, the segregation of observations (wafers) in the horizontal direction is more pronounced than in the vertical direction, and the horizontal direction separates out the chamber C wafers.

The loadings displayed in Figure 11.18 can be used to identify which process variables discriminate between the three chambers. In order not to reveal sensitive information, the variable points have been color coded according to which type of process factor category they represent. Further, the three artificial Y-variables are indicated by A, B and C. Two groups of variables have been marked by small ellipses. The left-hand group corresponds to a few variables reflecting ozone concentrations during the wafer manufacturing. These variables are systematically higher when processing is made with chamber C than with chambers A and B. Conversely, the right-hand group of variables denote variables that are systematically lower for chamber C. The latter group of variables represent equipment setting parameters.

The discussion above thus suggests how we can identify discriminating variables that are characteristic for the processing environment of chamber C. By knowing the identities of all the variables, one can, in a similar way, work out which variables are the discriminating variables for chambers A and B. Such detailed analysis is beyond the scope of the current chapter, however. In the next section we shall focus on the orthogonal components of the OPLS-DA model.

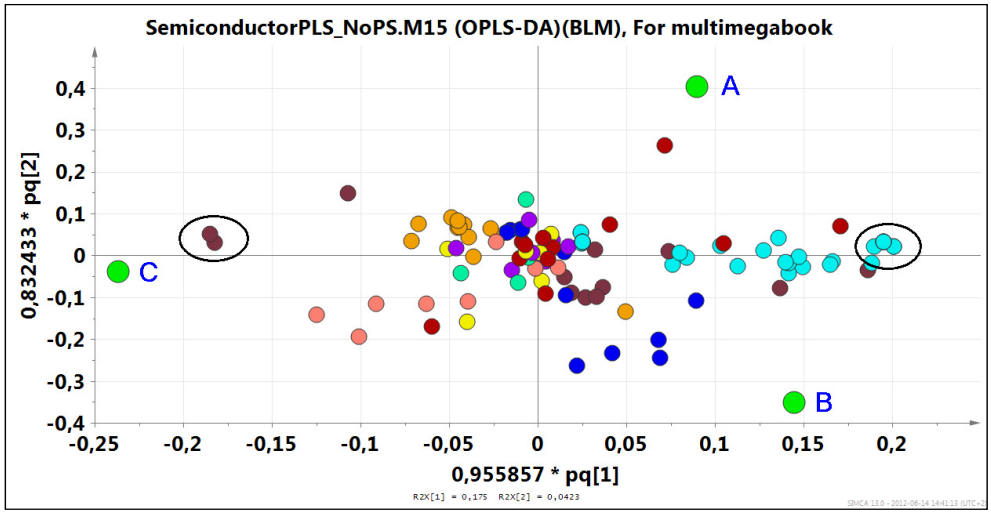


Figure 11.18: Loading plot corresponding to the score plot seen in Figure 11.16. The coloring of the variables is according to main type of process factor category. The location of the three artificial Y-variables is indicated by the capital letters A, B and C. Two groups of discriminating variables for chamber C have been marked by small ellipses (see text for discussion).

11.5.4 Interpretation of the orthogonal components

The OPLS-DA model obtained has three orthogonal components. In this section we interpret the two largest of these, which together account for close to 20% of the variation in the process variables (see Figure 11.15). The relevant score plot is graphed in Figure 11.19. As shown by Figure 11.19, there are four main clusters and one tiny cluster with less than ten points. The largest cluster is distributed near the origin. It contains the 1770 wafers that were well segregated by the two predictive components (see Figure 11.16). For these wafers little systematic variability remains beyond the two predictive components and hence it is logical that they are grouped near the center in Figure 11.19. The other three larger clusters correspond to groups of wafers for which there has been systematic structure to some process variables, but which is unrelated to the analysis question (to discriminate among the three chambers).

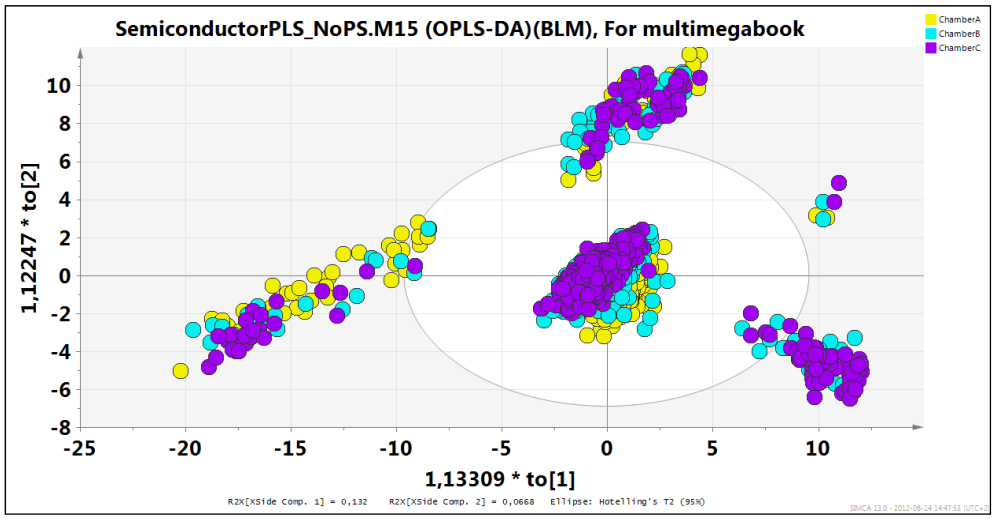


Figure 11.19: OPLS-DA score plot of the first two orthogonal-in-X components. Each point is one wafer. Coloring is according to process chamber (A, B or C).

Figure 11.20 displays the loadings of the first two orthogonal-in-X components. Again the variable points have been color coded according to which type of process factor category they represent. Three variables, each indicative of a major cluster of wafers in Figure 11.19,

have been marked by small ellipses. These variables and their connection to the clustering seen in the score plot of Figure 11.19 are elaborated on below.

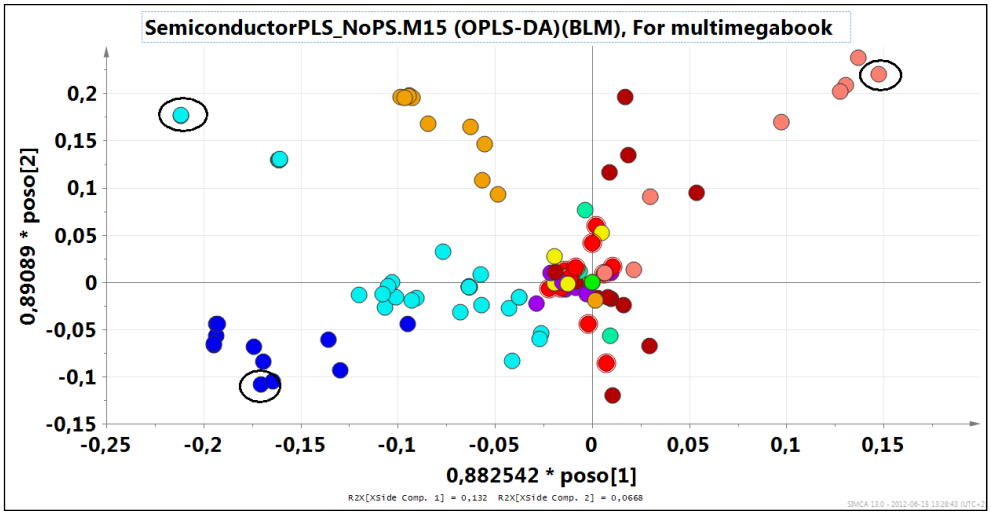


Figure 11.20: Loading plot corresponding to the score plot seen in Figure 11.19. The coloring of the variables is according to main type of process factor category. Three variables have been marked to help out in the model interpretation.

The sequence of plots shown in Figures 11.21 – 11.26 exemplifies how characteristic variables, one for each highlighted cluster of wafers, can be identified and interpreted. This connection between score plots, loading plots and raw data plots is very powerful in order to interpret the basis for any type of clustering detected among observations in a discriminant analysis model. Of course there is more than one variable behind the clustering investigated in Figures 11.21 – 11.26; our objective is simply to exemplify how orthogonal structure can be detected and interpreted.

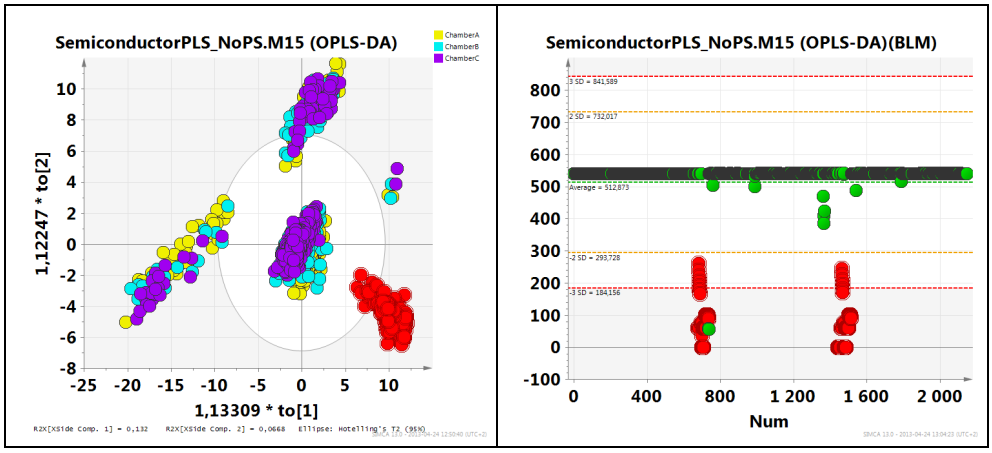


Figure 11.21: (left) The same score plot as in Figure 11.19 with one cluster highlighted (red color). This cluster contains 110 wafers. Figure 11.22: (right) Raw data plot of the variable marked in the top left corner of the loading plot of Figure 11.20. This is a temperature variable which is systematically lower for the wafers highlighted in Figure 11.21. This difference is unrelated to which chamber was used in the chip manufacturing and hence it appears in the orthogonal-in-X components.

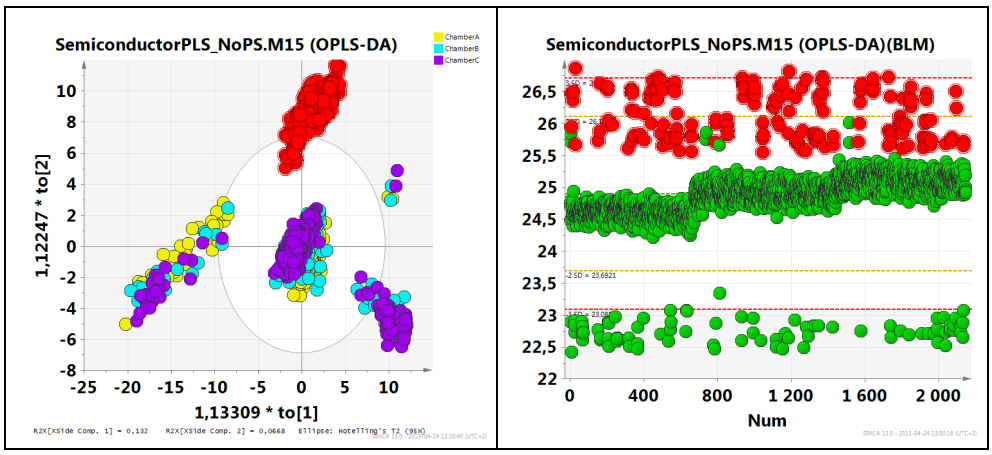


Figure 11.23: (left) The same score plot as in Figure 11.19 with one cluster highlighted (red color). This cluster contains 170 wafers. Figure 11.24: (right) Raw data plot of the variable marked in the top right corner of the loading plot of Figure 11.20. This is a variable reflecting the angle setting of a throttle valve, which is systematically higher for the wafers highlighted in Figure 11.23. This difference is unrelated to which chamber was used in the chip manufacturing and hence it appears in the orthogonal-in-X components.

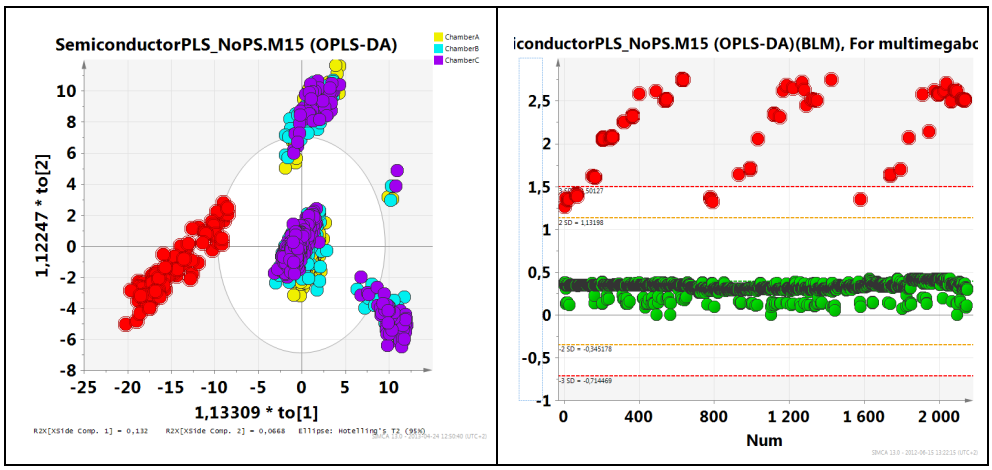


Figure 11.25: (left) The same score plot as in Figure 11.19 with one cluster highlighted (red color). This cluster contains 90 wafers. Figure 11.26: (right) Raw data plot of the variable marked in the lower left corner of the loading plot of Figure 11.20. This is a pressure variable, which is systematically higher for the wafers highlighted in Figure 11.25. This difference is unrelated to which chamber was used in the chip manufacturing and hence it appears in the orthogonal-in-X components.

11.5.5 Summary of the Wafer example

In contrast to the Poplar example, the Wafer example contains no obvious reference or golden standard that can be used as one uniting theme in a series of nested two-class models. On the contrary, the three classes (chambers) have similar independence and formulating the data analytical problem as a three-class discriminant problem is therefore a logical thing to do. The resulting OPLS-DA model was interpreted with special attention given to the three orthogonal components. The orthogonal components should not be overlooked. Understanding the background to such orthogonal variability aids in the total understanding of the manufacturing process and facilitates its long-term fine-tuning by removal of undesired or unfavorable systematic variability in the process factors.

11.6 Questions for Chapter 11

1. Why is the two-class problem preferable to the multi-class problem?
2. What is the S-plot? When is it applicable?
3. What is the S-line plot? When is it applicable?
4. What is the SUS-plot? When is it applicable?
5. What can the predictive components uncover in process data analysis?
6. What can the orthogonal components uncover in process data analysis?

11.7 Summary and discussion

OPLS-DA is a powerful method to identify discriminating variables. Whenever possible it is strongly recommended to formulate the discriminant problem as a two-class problem. This is because the resulting OPLS-DA model can then only have one predictive component and this is the easiest situation for revealing the best discriminating variables. It should be noted that the number of predictive components in an OPLS-DA model is always one less than the number of classes, i.e., with a three-class discriminant problem two predictive components will result, a four-class discriminant case gives three predictive components, and so on. The number of orthogonal components, however, will vary from case to case. One or two orthogonal components is not uncommon, but a much higher number of orthogonal components might indicate difficulties in appropriately setting up the classes that are being discriminated. A high number of orthogonal components indicate that there can be outliers, jumps, discontinuities or sub-classes within one or more of the existing classes. It is hence critical that the orthogonal components are interpreted in detail to understand what structure they represent. In fact, the orthogonal component in Figures 11.9 and 11.10 do indicate a subgroup formation within the wild-type class in the Poplar example. In the original work this subset formation was interpreted as related to structural changes in the xylem and phloem that make up the big transportation system of vascular plants.

When the discriminant problem involves three or more classes, care has to be exercised to avoid setting up ill-posed problems. When one of the classes is a reference class or a 'standard' class, the appropriate discriminant problem is the two-class problem. In biology and omics related problems, the reference class can be a control class or a wild-type class. When there is one control class and e.g. two treatment classes, such as in the Poplar dataset, it is not recommended to carry out the data analysis as a three-class discriminant problem. The reason for this is that one would then relate the three classes (control group plus the two genetically modified groups) to some hypothetical average class lacking relevance in reality. A similar misconception would be to specify a three-class discriminant problem involving e.g. a control group, a low-dose treatment group and a high-dose treatment group. Again, the center-point in the model would be a constellation of hypothetical observations lacking connection to reality.

On the contrary, however, the three-class problem and other multi-class problems are justifiable when there is no natural standard or reference class to compare with. This situation is exemplified in the last process example based on data from the semiconductor industry. Here, the three classes of data were created from three chambers supposedly being identical in their performance. The analysis of the Wafer dataset showed that one of the classes, the chamber C group, deviated most. Sometimes this may be a desirable phenomenon, sometimes not. In such a case, the outcome of the analysis may change the priority among the pre-defined classes. One may desire to define one of the classes to become the standard, and then there can be an interest of calibrating the rest of the classes against the emerging new reference class. If, for instance, it was known in the Wafer application that chamber C produced better goods, matching chambers A and B against it would be fully conceivable.

We also remind that the more classes that are included in a discriminant problem, the correspondingly more complex the resulting model will be to overview and interpret. We note that the specification of classes is part of the problem formulation. It must be done with caution and correspond to a real observable group similarity. The similarity within a class must be greater than the between-class similarity. Thus, an objective of the initial data analysis is to reveal whether a sound class division has been accomplished.

Furthermore, in this chapter we have introduced three plots geared for the identification of the most discriminating variables. These are the S-plot, the S-line plot and the SUS-plot. The most straightforward use of these involves the predictive components, but in case the orthogonal structure in X is large, assessing it by the S-plots is recommended.

Often it is desirable to ascribe a "number" to a discriminating variable in some way quantifying how good it is at separating the groups of observations being contrasted. In addition to loadings, coefficients and VIP, the literature is full of such variable discriminatory power measures. The book of Richard Brereton contains a very good introduction to many of the most commonly used class separation indices in pattern recognition [Brereton, 2009]. Another good reference is the paper written by Worley and coworkers [Worley, *et al.*, 2013].

Recently, a new measure for the importance of predictor variables, X, for the separation of two groups (classes) of observations was introduced [Eriksson and Wold, 2010]. This measure is a Graphical Index of Separation (GIOS), and is, for each predictor, determined from the distribution of all possible pairs of observations with one from each group. GIOS is quantitative, intuitively simple, and easy to interpret. The approach provides a ranking of predictor variables, X, including multivariate scores. The GIOS is straightforward to visualize in bivariate plots, and line or bar plots for larger numbers of variables. The approach applies both to discriminant analyses such as LDA, SIMCA, PLS-DA, OPLS-DA and to quantitative modeling such as MLR, PLS and OPLS. In the latter case, the observations are first divided into two groups based on their response values, Y.

12 Transformation and expansion

12.1 Objective

The objective of Chapter 12 is to address two data pre-treatment methods, which are useful for dealing with non-linear regression problems. The first method involves transformation of variables to achieve approximate normality, and the second, expansion of the X-matrix with higher-order terms.

12.2 Transformation

12.2.1 Introduction

Pre-treatment of measured data is carried out to re-shape ("transform") the data in order to facilitate data analysis and model interpretation. In Chapter 3, Section 3, we discussed the benefits of mean-centering and scaling to unit variance. However, data pre-treatment often has wider scope than just these two procedures. Scaling can be done in many different ways, depending on the modeling objectives and the level of prior knowledge about the properties of the data. Methods for scaling of data are further addressed in Chapter 13.

In spectroscopic applications, spectral data are often transformed in order to remove undesired systematic behavior ("signal correction"), or peculiarities that are unpleasant to the eye. Usually, such signal correction is carried out to improve the predictive power and interpretability of a multivariate calibration model, or to facilitate calibration transfer. Furthermore, when very large sets of spectral data are investigated, the pre-treatment phase may also involve measures to reduce the size of the data material ("signal compression"), for instance by using wavelet analysis. Tools for signal correction and signal compression are discussed and exemplified in Chapter 14.

Another reason for transforming raw data occurs when a variable contains one or a few extreme measurements, which may influence model building unduly. Consider Figure 12.1, which shows the histogram of a variable called Var1. One out of the 40 measurements in this variable is substantially larger than the others. If this extreme measurement is not manipulated in some way prior to data analysis, it will exert a large influence on the model and dominate over the other measurements.

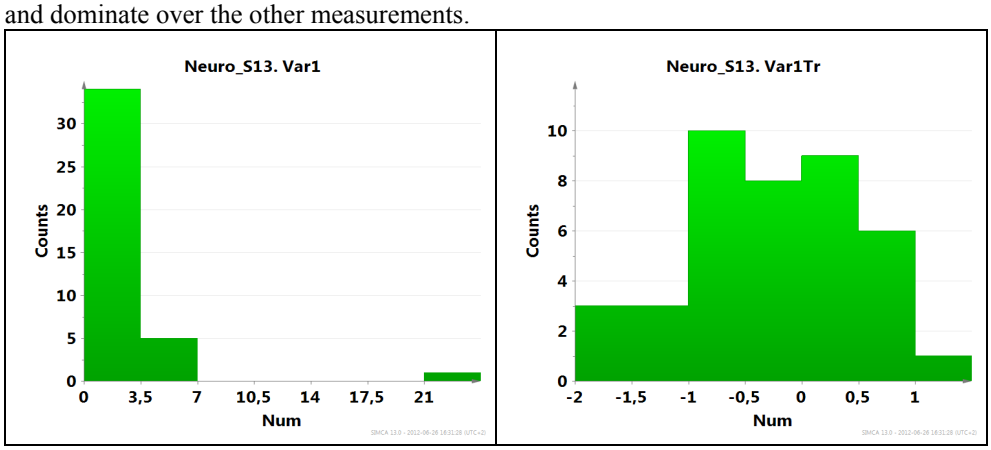


Figure 12.1: (left) Histogram of a non-transformed variable Var1. Figure 12.2: (right) Histogram of variable Var1 after log-transformation.

This problem can be formulated in another way by stating that the distribution of Var1 departs from normality, that is, it has a positive skewness. This deviation from a normal distribution means that we will be unable to achieve high efficiency in the data analysis. Hence, an objective of data pre-treatment might in this case be to convert the non-normal distribution of Var1 into a normal one. One way to make a non-normal distribution more nearly normal is by *transformation*, for example, the log-transformation [Davies & Goldsmith, 1986; Massart, *et al.*, 1988]. Figure 12.2 displays the result of log-transforming Var1. Apparently, Var1 is approximately normally distributed after log-transformation. It is noted that a distribution of log (x) will not necessarily be perfectly normal, but will usually be much closer to normality than is the case for the untransformed data.

As evidenced by Figure 12.1, the histogram plot is an appropriate tool for discovering the need for transformation in order to approach normality. However, there are other tools available, as well. In Section 4.7.1 it was demonstrated that a PLS score plot of the t/u-type is useful for identifying cases where either responses or factors need to be transformed. Also note that sometimes data may contain very extreme observations ("outliers"), which may not be addressed satisfactorily by transformation. Then, other approaches such as trimming or winsorizing may have to be tested (see discussion in Section 13.6).

The remaining part of Section 12.2 will be devoted to a discussion of a set of transformations that are useful for achieving normally or approximately normally distributed data, and some benefits of such transformations.

12.2.2 Which transformations are useful for approaching normality?

The log-transformation is not the only transformation one can think of. Other often used transformations, which are supported by SIMCA, are negative logarithm ("neglog"), logit, square root, fourth root, inverse, and power transformations (see Appendix II). There is no doubt, however, that the log-transform is the most frequently applied transformation. This is because log-normal distributions are often encountered in nature, particularly when the variable studied has a *natural zero*, such as, retention times, weight, height, concentrations, etc., and ranges over one order of magnitude. Regardless of which transformation is selected, a histogram plot of each variable is a good way of detecting deviations from normality. Figures 12.3 to 12.5 show three typical outcomes of histogram plotting, where log-transformation, no transformation, and neglog-transformation are applicable.

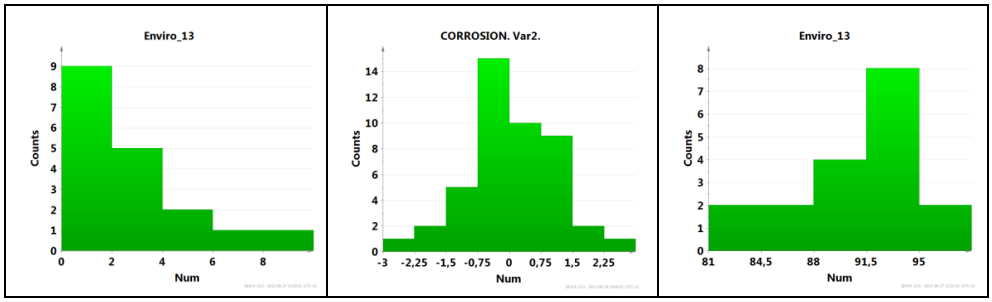


Figure 12.3: (left) Histogram of a variable where the log-transformation is applicable. Figure 12.4: (middle) Histogram pointing to a variable where no transformation of the raw data is needed. Figure 12.5: (right) Histogram of a variable where the distribution is skewed to the left, and hence the negative logarithmic transformation is warranted.

Sometimes it is possible to extract clues for transformation by considering the nature of the variable in question. Suppose, that we are working in the car industry. Further assume that our task is to detect a good car paint, which will make the car as resistant to corrosion as possible. Then we might explore this resistance by trying many different car paints. After exposure to "real climate conditions", we might quantify the resistance to corrosion by measuring the size of occurring rust spots, say, in mm^2 or cm^2. For most tested objects we would hopefully score near-zero values, but certainly there would be a few cases where large rust spots would be registered. If plotted in a histogram, this kind of size variable

would display a positive skewness (cf. Figure 12.3), thereby indicating a non-normal situation. One method to approach normality could be to employ the square root transform, as this transform would be logically linked to the manner in which the response is expressed.

12.2.3 Benefits of transformation

We will now study some benefits of variable transformation. In regression modeling, a properly selected transformation of a response variable brings about a number of benefits. Notably, it may (i) simplify the response function by linearizing a non-linear response-factor relationship, (ii) stabilize the variance of the residuals, and (iii) make the distribution of the residuals more normal, which effectively implies that outliers are eliminated. As an illustration, we shall consider a DOE application aimed at producing a long-lasting device for service in an aircraft construction. Ten factors (X-variables) were varied using a screening design consisting of 32 experiments (observations), and the measured response was the lifetime in hours. The design employed supports a linear model.

As displayed in Figure 12.6, when fitting the linear model to the data, the statistics R^2Y = 0.88 and Q^2Y = 0.60 were obtained. These results are reasonably satisfactory. However, a closer scrutiny of the model in terms of some of the residual plots strongly suggests that something is not right with the computed model. Figures 12.7 – 12.9 provide plots displaying some of the modeling problems, problems that hopefully might be relieved with a suitable response transformation. In Figure 12.7, the observed and estimated response values are plotted against each other. From the strong curvature observed, we can conclude that the response function between the ten factors and the lifetime response is non-linear.

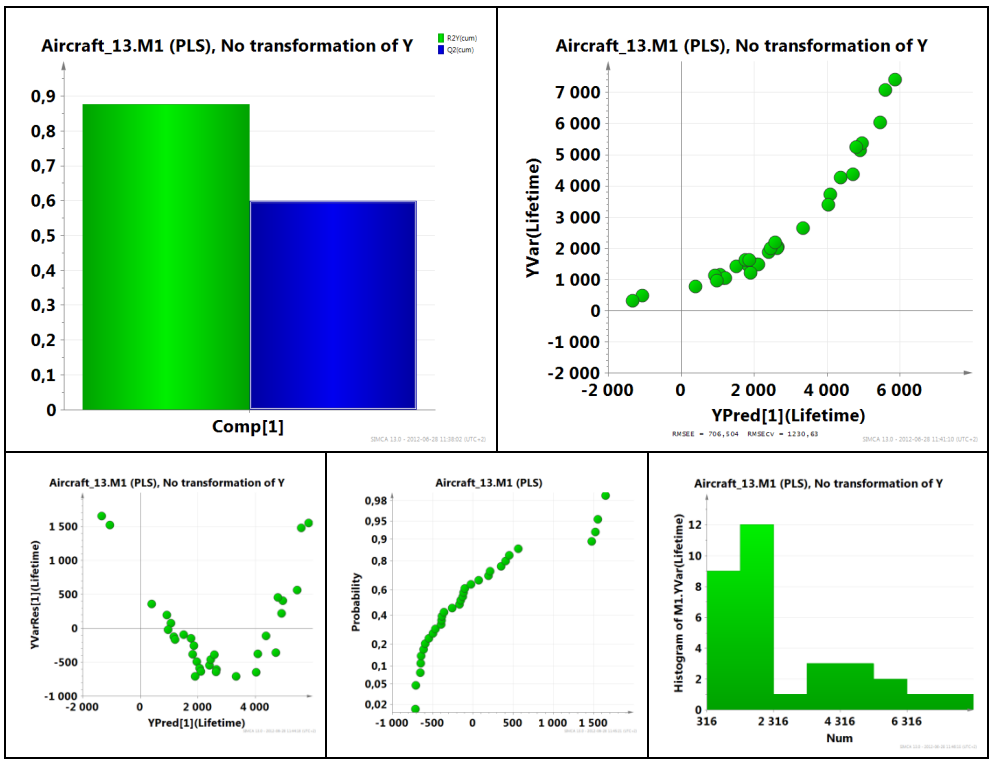


Figure 12.6: (upper left) Summary of fit plot – before transformation. Figure 12.7: (upper right) Observed and estimated response data – before transformation. Figure 12.8: (lower left) Residuals versus predicted data – before transformation. Figure 12.9: (lower middle) N-plot of residuals – before transformation. Figure 12.10: (lower right) Histogram of response – before transformation.

Further, in Figure 12.8 the modeling residuals are plotted against the estimated response values. In this kind of plot one does not want to see any systematic structure. In our case, however, a strong boomerang-like data distribution is seen, suggesting some structure to the variance of the residuals. This is unwanted. Similarly, in Figure 12.9, the normal probability

plot of residuals shows a structure in the residuals, i.e., a group of four deviating experiments. In this plot, the desired result is that all points are situated on a straight line which goes through the point (0, 50%). Because of the non-linear appearance of the residuals in the N-plot, it is logical to anticipate that the response is unsymmetrically distributed. That this is indeed the case is evidenced by the histogram displayed in Figure 12.10.

Consequently, the lifetime response was log-transformed and the linear model refitted, the results of which are illustrated in Figures 12.11 – 12.15. Figure 12.11 shows that R^2Y now becomes 0.99 and Q^2Y 0.77, which are significant improvements justifying the transformation undertaken. In addition, the response transformation has (i) simplified the response function by making the response-factor relationship linear (Figure 12.12), (ii) made the size of the residual independent of the estimated response value (Figure 12.13), and (iii) made the distribution of the residuals more nearly normal (Figure 12.14). The histogram of the log-transformed response is plotted in Figure 12.15 and it shows that the response is much more symmetrically distributed after the transformation.

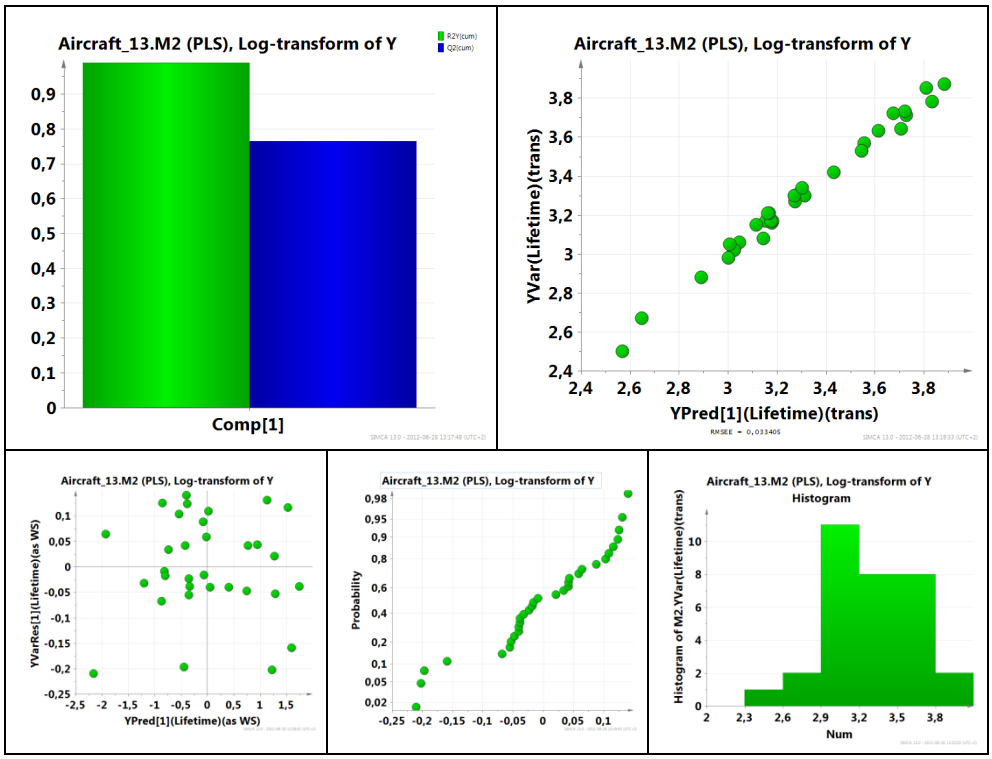


Figure 12.11: (upper left) Summary of fit plot – after transformation. Figure 12.12: (upper right) Observed versus estimated response data – after transformation. Figure 12.13: (lower left) Residuals versus predicted data – after transformation. Figure 12.14: (lower middle) N-plot of residuals – after transformation. Figure 12.15: (lower right) Histogram of response – after transformation.

In summary, the need for transforming a response is frequently detected in a histogram plot, but may be found from many other kinds of plot, as well. And, as seen in the outlined example, a carefully selected transformation may make the regression modeling simpler and more reliable.

12.3 Example: NEURO

We will now study a dataset where transformation of variables will affect a PCA model profoundly. The dataset contains 40 observations, which are neuroleptic compounds. Such compounds exhibit a wide variety of pharmacological activities. In order to map their pharmacological potencies, a test battery of 12 pharmacological endpoints in rat was used. These variables express the potency of each compound as ED_{50} in mg/kg. The dataset was

re-published and analyzed by multivariate analysis by Lewi [Lewi, 1995]. One objective of the study was to examine whether the 40 substances could be classified according to activity profile.

PCA was used to investigate the 40 (observations) by 12 (variables) data table. The first two components were calculated for overview purposes, but note that only the first one was significant according to cross-validation. After two components $R^2X = 0.66$ and $Q^2X = 0.36$ (Figure 12.16). The t_1/t_2 score plot reveals that some compounds deviate substantially from the majority of neuroleptics (Figure 12.17).

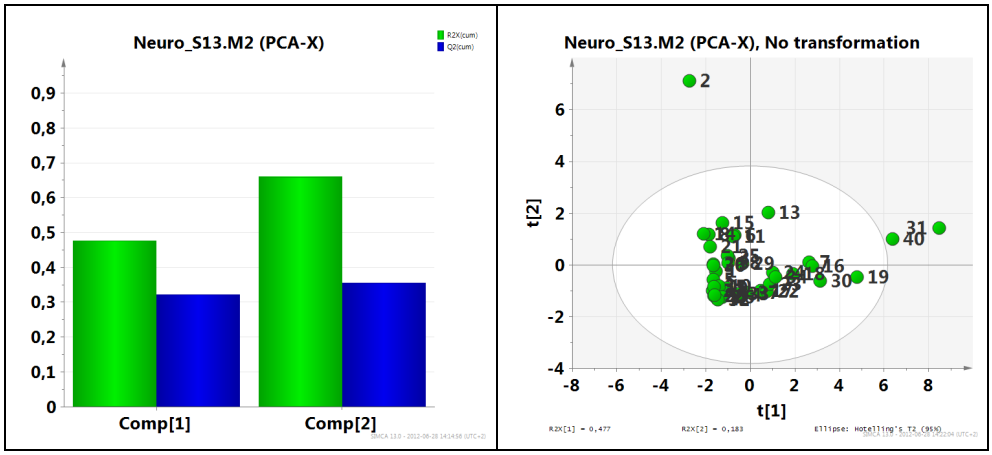


Figure 12.16: (left) Summary of fit of original PCA model of NEURO dataset. Figure 12.17: (right) PCA t_1/t_2 score plot of model for NEURO dataset. Each compound is shown as a unique point. Some compounds deviate substantially.

As was pointed out in Chapter 9, the contribution plot is a useful tool for understanding what makes a certain observation different from another observation, or the average observation. Figure 12.18 shows a contribution plot between observation 31 and the average observation. We can see that observation 31 has higher numerical values in variables 1-3, and 5-7 than the average observation. In an analogous fashion, Figure 12.19 indicates that observation 2 differs from the average observation because of higher values in variables 8, 9, and 12.

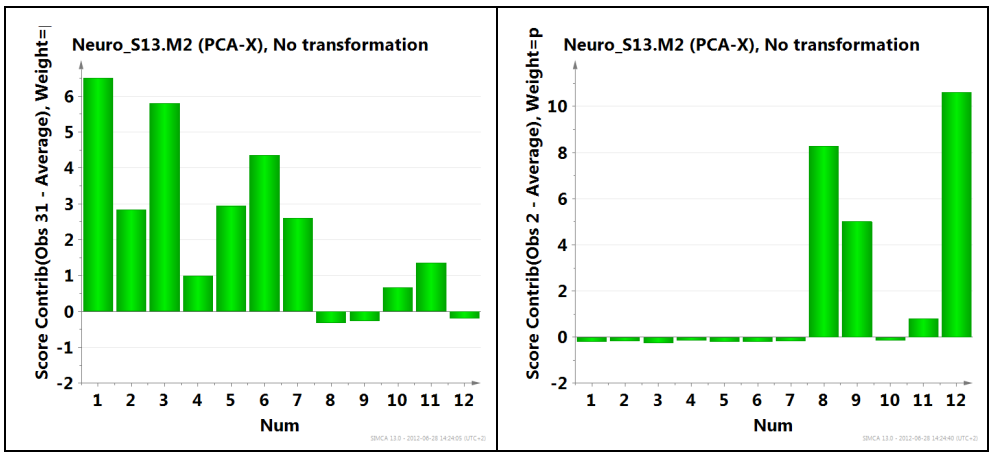


Figure 12.18: (left) Contribution plot (scores mode) between observation 31 and the average observation.
Figure 12.19: (right) Contribution plot (scores mode) between observation 2 and the average observation.

The contribution patterns seen in Figures 12.18 and 12.19 make it of interest to examine the distributions of the individual variables, simply to see whether more observations have extreme values in their descriptors. Instead of producing one histogram for each variable, a convenient alternative for variable overview is the skewness of each variable. This parameter measures the lack of symmetry of a distribution (see Appendix II).

Figure 12.20 gives the skewness of each variable together with the corresponding min/max ratio. A near-zero value of skewness indicates a symmetrical distribution of raw data (cf. Figure 12.4). A positive skewness corresponds to a distribution tailing to the right (cf. Figure 12.3), and a negative value to a distribution with a tail to the left-hand side (cf. Figure 12.5). In order to obtain a suitable test statistic for skewness, the calculated skewness value is divided by the corresponding standard error. More details are found in Appendix II.

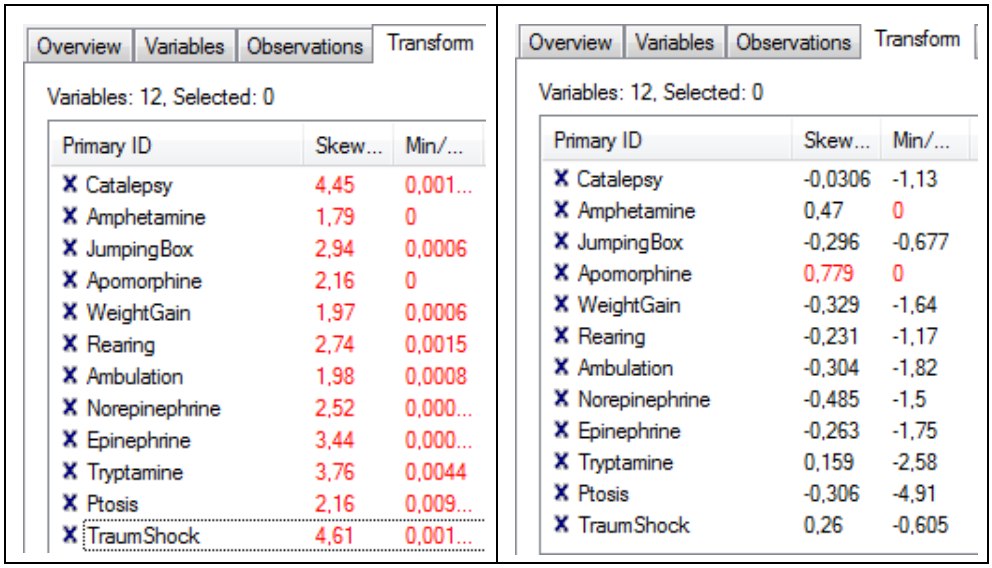


Figure 12.20: (left) Min/max ratio and skewness of each variable prior to transformation. Figure 12.21: (right) Same statistics as in Figure 12.20, but after log-transformation.

SIMCA uses a combination of the skewness test statistic and the min/max ratio to uncover which variables are candidates for a transformation. In the NEURO dataset, all twelve variables might benefit from a transformation. Because of the nature of the data (biological activities with a natural zero) we selected the log-transform. Variables 2 and 4 were transformed after adding a small constant (+1), while variables 1, 3, and 5-12 were transformed without adding any constant. Figure 12.21 shows min/max ratio and skewness after taking the logarithm of each variable. Evidently, skewness is now closer to 0 in all cases.

The PCA modeling of the log-transformed X-matrix yielded two components with $R^2X = 0.81$ and $Q^2X = 0.71$ (Figure 12.22). Figure 12.23 provides the scores of this model. It is clear that the distribution of compounds is much more uniform in the score space. Hence, the transformations used have been successful in removing some of the extreme behavior previously exhibited by compounds 2 and 31. The rather homogeneous spread of the observations indicates that the 40 compounds are not strongly clustered according to pharmacological activity profiles.

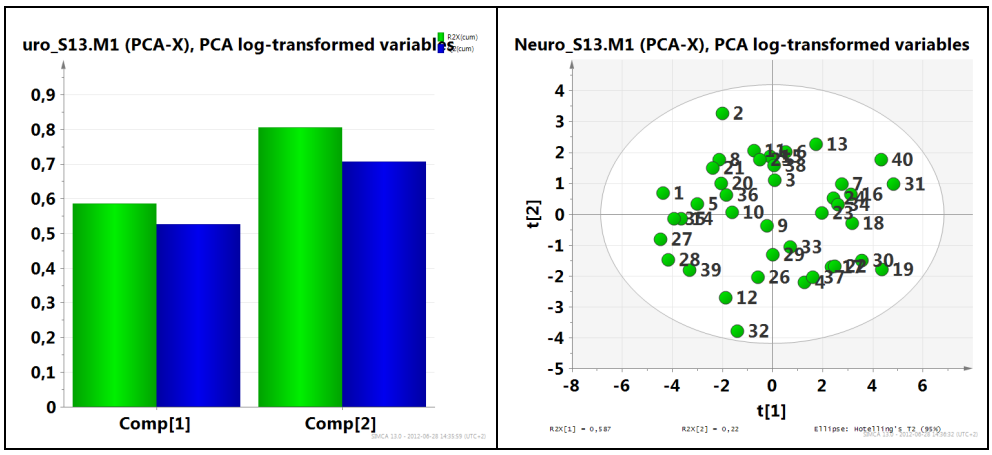


Figure 12.22: (left) Summary of fit of new PCA model of log-transformed NEURO dataset. Figure 12.23: (right) PCA t_1/t_2 score plot of new model for NEURO dataset. The rather homogeneous distribution of compounds indicates that no strong grouping according to pharmacological profiles exists.

In summary, the NEURO application shows that by carefully selecting a variable transformation, data may be made more approximately normal. This, in turn, will enhance the efficiency of the data analysis.

12.4 Expansion of the X-matrix

12.4.1 Introduction

The central idea in design of experiments (DOE) is to perform representative and diverse experiments, by manipulating all relevant factors at the same time. All variables ("factors") in the X-matrix that are varied according to DOE will become mathematically independent ("orthogonal"). This makes it possible to expand the X-matrix with complementary model terms (Figure 12.24) like interaction ("cross") and quadratic ("square") terms, which are able to capture non-linearities between X and Y. The complexity of the expanded terms depends strongly on the experimental protocol selected. This is further discussed in Chapter 18.

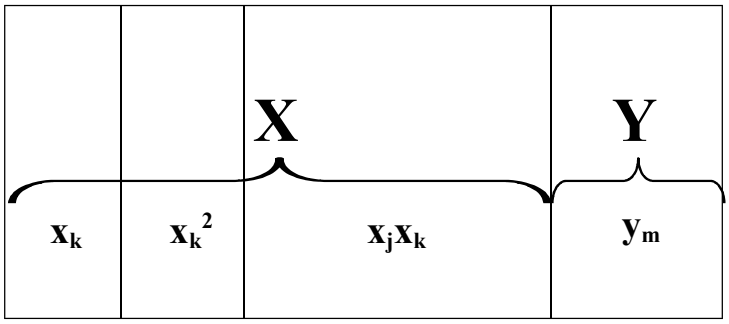


Figure 12.24: Expansion of the X-matrix with interaction terms ($x_j x_k$) and quadratic terms (x_k^2) is possible when an experimental design of the central composite family is utilized. The expanded terms are used alongside the parent terms in the regression modeling.

Expanding the X-matrix with interaction and quadratic terms is not a problem when DOE has been implemented. However, there may be problems with non-orthogonal datasets. The first problem concerns the foundation of the interaction terms. Because many of the original X-variables usually are correlated, the estimation of interaction terms among them will be unreliable. There is often no experimental support for interaction terms.

A second and even greater problem is related to redundancy. For example, an expansion of 50 X-variables with quadratic and interaction terms will result in 50 (linear) + 50 (quadratic)

+ 50*49/2 (interaction) = 1325 X-variables. It will certainly not be the case that all these variables are meaningful and of appreciable size. This large number of additional variables will dominate the model and add noise to the X-block, often resulting in poor predictive performance.

12.4.2 Implicit non-linear latent variable regression, INLR

Interestingly, however, Berglund and co-authors have shown that for certain types of problem it is not necessary to include interaction terms [Berglund & Wold, 1997a]. Rather, it is sufficient to include only quadratic terms of the X-variables. This approach is called *implicit non-linear latent variable regression* (INLR). Its underlying assumption is that the X-matrix has a *latent structure* which correlates with Y, and that this structure is modelable by computing latent variables summarizing the original X-variables. With a non-linear relationship between Y and the latent structure of X, the idea is that an expansion of the latent variables will be sufficient for obtaining an adequate regression model. This means that it is crucial that there be enough X-variables to support the estimation of the latent structure, and its increased dimensionality following the expansion. INLR has been shown to work well in calibration and QSAR problems, where X has a latent structure and the rank is much lower than the number of variables measured. INLR and other tools for non-linear regression modeling are further addressed in Chapter 20.

12.5 Example: SOVRING

The SOVRING application, which was introduced in Section 9.6, is an illustration of how expansion of the X-matrix may be carried out in a real process environment. We recall that three out of the twelve X-variables were varied according to a 17 run experimental design. The design used supports a quadratic model in the three designed factors (Figure 12.25). In this short account, we will only briefly review the model computed on the designed X-variables (cf. Section 9.6.6). Attention is given to two of the six response variables, %P_FAR and FAR. The former displays a non-linear relationship with regard to the factor HS_2, whereas the latter is principally modeled by linear model terms.

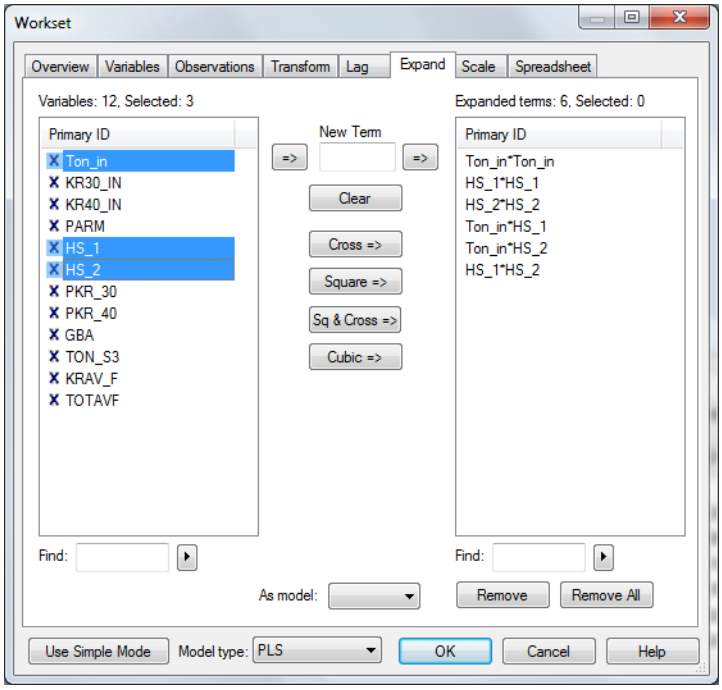


Figure 12.25: Illustration of matrix expansion of the X-matrix of the SOVRING dataset. The three original X-variables Ton_in, HS_1, and HS_2, are expanded by forming their interaction and quadratic terms.

Figure 12.26 displays the regression coefficients of the response %P_FAR. We can see that the speed of the first magnetic separator (HS_1) has limited impact on %P_FAR. In order to

decrease the amount of phosphorous in the FAR concentrate, one should focus on running on a low load and use high speed of the second magnetic separator. The quadratic term of HS_2 contributes significantly to the modeling of %P_FAR. Figure 12.27 displays a response contour plot of %P_FAR, where the strong curvature arises from the importance of the HS_2 quadratic term.

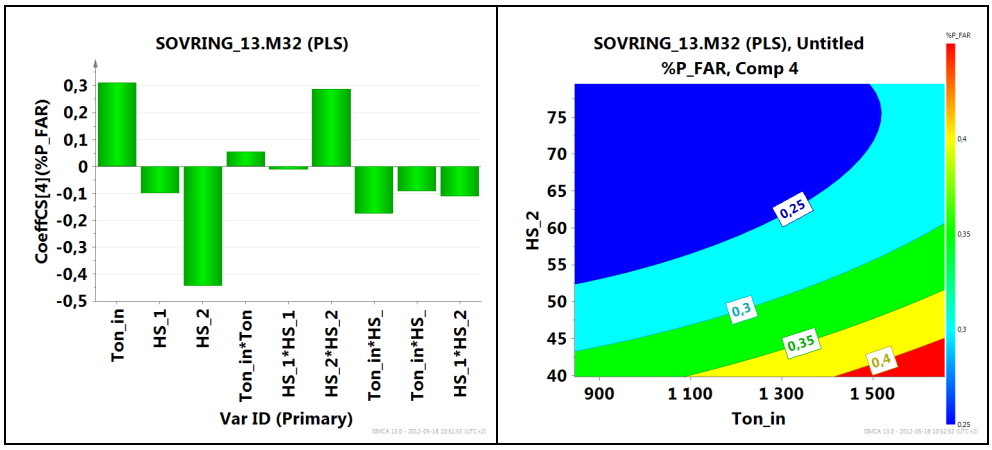


Figure 12.26: (left) Regression coefficients of %P_FAR. The PLS model is fitted to the designed factors only. The factors Ton_In and HS_2 have the highest impact on the response. There is a significant quadratic behavior with respect to HS_2. Figure 12.27: (right) Response contour plot of %P_FAR, where the influences of Ton_In and HS_2 are seen. The strong curvature of the response contour plot is modeled by the important quadratic term.

A completely different behavior is demonstrated by the response FAR, which is not as dependent on quadratic model terms (Figure 12.28). It is, in fact, mostly influenced by the linear term of Ton_In, that is, the more starting material coming into the plant, the more FAR is produced. Since the contribution from the quadratic model terms is limited, the response contour plot is less curved than in the previous case (Figure 12.29). In order to increase the amount of FAR, we should stick to the lower right-hand corner.

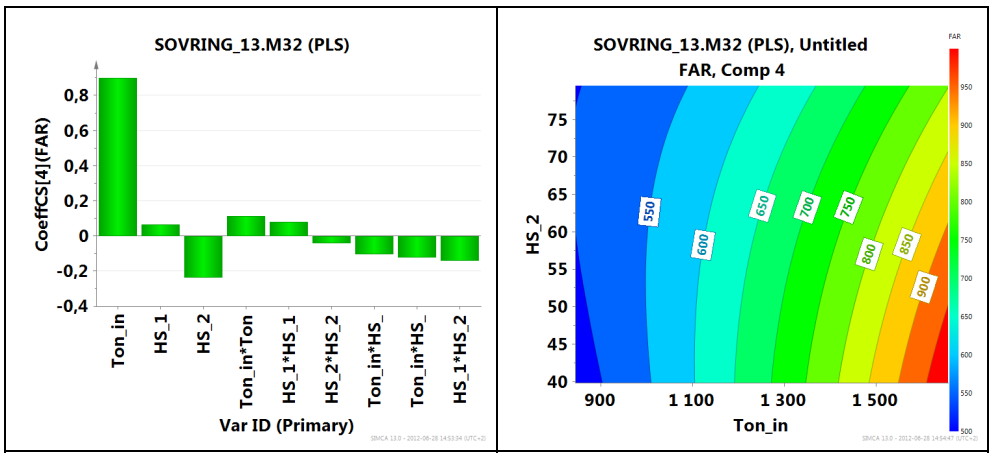


Figure 12.28: (left) Regression coefficients of FAR. The PLS model is fitted to the designed factors only. The factor Ton_In dominates the model, and in order to increase the amount of FAR Ton_In should be maximized. Figure 12.29: (right) Response contour plot of FAR, where the influences of Ton_In and HS_2 are seen. The absence of a strong quadratic influence makes the response contour plot much less curved than in Figure 12.27.

In summary, the ability to expand an X-matrix with interaction (cross) and quadratic (square) terms renders it possible to accomplish non-linear regression modeling. When PLS is used, this approach is often known as PLS-RSM (response surface modeling). In the SOVRING application, there is no doubt that some of the responses benefit from the inclusion of higher-order terms. This would not have been possible without the underlying RSM design. Design of experiments in process modeling and optimization is discussed in Chapter 18.

12.6 Questions for Chapter 12

1. Which transformations are useful for achieving approximately normally distributed data?
2. What are the benefits of transforming a response variable in regression modeling?
3. Which diagnostic tools are available for detecting the need for transformation?
4. What is the principal condition that needs to be fulfilled if an X-matrix is to be expanded with interaction and quadratic terms?
5. When is INLR useful?

12.7 Summary and discussion

In this chapter we have considered (i) transformation of variables and (ii) expansion of the X-matrix. Much of the discussion was focused on the regression situation. The primary incentive for transforming a (response) variable is that of maintaining a high efficiency of the data analysis. In addition, a correctly chosen transformation of a response variable may sometimes linearize a non-linear problem and thereby simplify the modeling. Furthermore, when an RSM design has been used for studying a system or a process, it is possible to expand the X-matrix with interactions and quadratic terms. Such terms make it possible to model non-linearities between factors (X) and responses (Y). Moreover, in cases where non-designed X-data contains a strong latent structure, a selective expansion with quadratic terms of the measured variables has been found useful for modeling non-linearities. This latter approach is called INLR.

13 Centering and Scaling

13.1 Objective

In Chapter 13 the objective is to outline principles for the scaling of data. Scaling is an integral part of data pre-processing, and is a step that must be carried out cautiously. We will mainly concentrate on *no scaling*, *Pareto scaling*, and *scaling to unit variance*, but some related methods will be considered, as well. This chapter also provides a discussion regarding *Trimming* and *Winsorizing*.

13.2 Introduction

The pre-processing phase of multivariate data involves a number of steps. Trimming and winsorizing are optional first steps to remove gross errors and influential outliers from the dataset [Wold, *et al.*, 2002; Kettaneh, *et al.*, 2005], see further discussion in Section 13.6. As mentioned in Section 12.2, the data pre-processing, in turn and order, involves tools for transformation (Chapter 12), centering & scaling (Chapter 13) and signal correction & compression (Chapter 14). Scaling is the main topic of this chapter.

For many types of data, centering and scaling are intertwined. Centering, which was introduced in Section 3.3, corresponds to a subtraction of a reference vector (often represented by the mean values of the variables or the settings of the set point). Scaling corresponds to a multiplication by a vector. The choice of scaling vector is crucial. In many areas, such as QSAR, multivariate design, sensory analysis, etc., where variables of different origin and numerical range are encountered, the scaling vector is chosen as the inverse spread (reciprocal standard deviation) of the variables. In other situations, for example with process data, the scaling vector is defined relative to a tolerable spread in the variables. Some of these cases are discussed below, but first we give a few general remarks about scaling as a concept.

A *scale* is a representation of a number of observations on a line such that the positioning of these points is related in some mathematical sense to one or more criteria relevant to these observations. The real numbers that are used to formulate a scale may be characterized by the attributes *order, distance*, and *origin* [Jackson, 1991]. Firstly, the numbers are ordered. This implies that an observation with a numerical value of 13 is higher in some respect than an observation which has a value of 5. Secondly, differences (distance) between numbers are ordered. This means that two observations with scale values of 10 and 20 are farther apart than two other observations having numerical values of 5 and 10. Thirdly, the series of real numbers has a unique origin, the number zero. Based on these attributes, scales may be described as either *ordinal*, *interval*, or *ratio* scales, depending on whether the first, the first and second, or all three attributes are present.

13.2.1 Scaling to unit variance (UV-scaling)

Techniques for *scaling* thus involve mathematical operations that modify the numerical values of a scale. Since PCA, PLS, OPLS and O2PLS are sensitive to the choice of scaling technique, it is important to carry out this step with care. We understood in Section 3.3 that if no prior information is available, scaling to unit variance (UV) is the common practice. Unit variance scaling implies that "long" variables are shrunk and "short" variables are stretched, so that all variables will rest on an equal footing, that is, unit variance. The scaling weight employed in UV-scaling is $1/s_k$, where s_k represents the standard deviation of

variable *k*. In addition to UV-scaling, mean-centering is often used, a combination often denoted *auto-scaling*.

If prior information about the importance of variables is available, this insight should be reflected in the scaling. Variables that are known to be "important" for a given problem can be up-weighted, to at least partially account for this relevance. Analogously, variables that are noisy, or known to be irrelevant for the problem at hand, can be assigned a lower weight to diminish their contribution to the modeling.

In general, however, if there is no prior information available, UV-scaling is the most objective way to scale the data. This will give all variables an equal opportunity of influencing the data analysis. UV-scaling is useful when variables are of different kinds and not directly comparable numerically.

13.2.2 No scaling and Pareto scaling

Sometimes, however, no scaling (but mean-centering) is the desired method for "scaling" the data. Usually, this option is deployed when all variables are expressed in the same unit, such as with spectroscopic data. Moreover, an alternative technique called Pareto scaling has become more common [Wold, 1993a]. Pareto scaling gives each variable a variance numerically equal to its initial standard deviation instead of unit variance. Here, the scaling weight is $1/\sqrt{s_k}$. Hence, Pareto scaling is intermediate between the extremes of no scaling and UV-scaling.

13.2.3 Set-point centering and control/action-limit scaling

In process modeling, a useful alternative to mean-centering and UV-scaling involves centering with regard to the set-point of controlled process variables and scaling according to the limits defined as control and/or action limits of the same variables. Set-point centering allows the process operator to focus on the variability around the set point and not around the average point. Control or action limit scaling may produce a more realistic model of the process, especially in the case when variables display a variation in the historical database that is much smaller than their normal or tolerated variation range.

13.2.4 Inclusion/exclusion of variables is a way of scaling data

Furthermore, from a philosophical point of view, the most significant scaling of data is conducted in the phase where the variables (X and Y) are selected. In principle, the *inclusion* of a variable in a dataset corresponds to assigning this particular variable an initial weight of one (1). This weight may later be altered as part of the data analysis. Similarly, the *exclusion* of a variable from consideration is equivalent to the assignment of zero (0) initial weight. This latter way of weighting, however, is decisive in nature, and often means that that variable will never be evaluated.

We will now illustrate the procedures of no scaling, UV-scaling, and Pareto scaling using the LOWARP and SUGAR datasets. Subsequently, a discussion regarding three further, related scaling procedures will take place.

13.3 Example: LOWARP

In the LOWARP dataset, the variables have markedly different numerical ranges, particularly among the 14 warp and strength responses (Table 13.1). The initial standard deviation of a variable measures its numerical variation. For variables st2, st3, st4, and st5, the initial standard deviation is in general 10 - 100 times larger than for any other variable. Thus, unless variables st2 - st5 are standardized by some type of scaling, they will dominate in all modeling attempts.

Table 13.1: The LOWARP dataset.

		1	2	3	4	5	6	7	8	9	10	13	16	11	12	14	15	17	18
Num	Name	glas	crtp	mica	amtp	wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	wrp7	wrp8	st1	st2	st3	st4	st5	st6
1	1	40	10	10	40	0.9	5	0.2	1	0.3	4.2	1.2	1.3	232	15120	2190	26390	2400	0.7
2	2	20	20	0	60	3.7	7.3	0.7	1.8	2.5	5.4	1.8	2.1	150	12230	905	20270	1020	0.6
3	3	40	20	0	40	3.6	6.9	0.9	2.1	4.8	9.4	1.2	1.4	243	15550	1740	21180	1640	
4	4	20	20	20	40	0.6	3.1	0.3	0.4	0.4	1.1	1	1	188	11080	1700	17630	1860	0.5
5	5	20	10	20	50	0.3	2.1	0.3	0.3	0.8	1.1	1.2	1.3	172	11960	1810	21070	1970	0.5
6	6	40	0	20	40	1.2	5					1.1	1.3	245	15600	2590	25310	2490	0.6
7	7	20	0	20	60	2.3	3.9	0.3	0.4	0.7	1.4	1.5	1.6	242	13900	1890	21370	1780	
8	8	40	0	10	50	2.6	5.9	0.4	0.2	0.7	1.2	1.6	1.6	243	17290	2130	30530	2320	0.7
9	9	30	20	10	40	2.2	5.3	0.2	0.7	0.6	2	1	1.1	204	11170	1670	19070	1890	0.6
10	10	40	0	0	60	5.8	7	0.9	1	5.6	11.8	1.6	1.8	262	20160	1930	29830	1890	
11	11	30	0	20	50	0.8	2.9	0.5	0.6	1.1	2	1.3	1.3	225	14140	2140	22850	2110	0.7
12	12	30	10	0	60	2.8	5.1	1	1.2	2.7	6.1	1.9	2.1	184	15170	1230	23400	1250	0.6
13	13	30	10	10	50	1.1	4.7	0.6	0.9	1.3	3.5	1.4	1.4	198	13420	1750	23790	1930	0.7
14	14	30	10	10	50	1.9	4.7	1	1	2.8	5.4	1.5	1.6	234	16970	1920	25010	1790	0.7
15	15	30	10	10	50	2.9	5.9	0.5	0.6	1	6.6	1.5	1.6	239	15480	1800	23140	1730	
16	16	40	10	0	50	5.5	7.9	0.8	2.4	5.5	9.3	1.5	1.8	256	18870	1880	28440	1790	
17	17	30	0	10	60	3.2	6	0.3	0.5	1.5	5.2	1.5	1.7	249	16310	1860	24710	1780	

The application of PLS to the UV-scaled and mean-centered data was described in Section 4.5.6. A two-component model was obtained, and the PLS weights are plotted in Figure 13.1. We can see that all variables, the four X-factors and the 14 Y-responses, participate in the formation of the regression model. An interpretation of this model was presented in Section 4.5.6.

A new PLS model was fitted to mean-centered data (i.e., with no scaling). As shown by Figure 13.2, this model is very different from the previous one. The PLS weight plot has a completely different appearance. The new model is dominated by the four responses st2, st3, st4, and st5, that is, the variables displaying the largest initial numerical variation. Hence, in the LOWARP case, scaling to unit variance is favorable compared with no scaling. The former approach increases model interpretability.

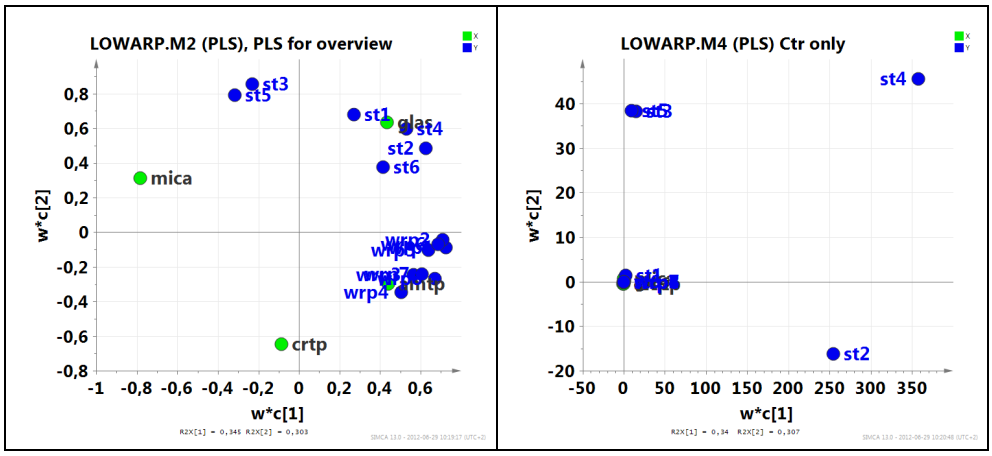


Figure 13.1: (left) PLS weight plot of model for UV-scaled and mean-centered LOWARP data. All 18 variables contribute to the PLS model. Figure 13.2: (right) PLS weight plot of model for unscaled but mean-centered LOWARP data. Now, only four variables, st2 – st5, contribute significantly to the model. These are the variables with the largest initial variations.

13.4 Example: SUGAR

In the LOWARP dataset, the choice of scaling technique was quite obvious. However, as we shall see in the SUGAR dataset, selecting the best way of scaling is not always so straightforward. In this discussion, we will not address the complete SUGAR dataset, but rather the reduced one with 90 observations sorted according to decreasing ash content (cf. Section 8.4.4). The division of the 90 observations into odd- and even-numbered sub-sets was preserved.

In the PLS modeling, three different scaling techniques were tested, viz., UV-scaling, Pareto scaling, and no scaling. Mean-centering was conducted for each. A PLS model was trained on one sub-set of observations and tested predictively on the other. This gave six combinations (PLS models), i.e., three scaling procedures times two sub-sets for model establishment.

The results of the PLS modeling efforts are given in Table 13.2. This table lists the explained variation of the X-block (R^2X), the explained Y- (ash content) variation (R^2Y), the predicted Y-variation according to internal cross-validation (Q^2_{int}), and the predicted Y-variation of the external prediction set (Q^2_{ext}). According to the R^2- and Q^2-estimates, the computed models are similar, at least as far as modeling and predictive powers are concerned. The largest difference relates to the X-block explanatory power, where UV-scaling utilizes a little less of X for predicting ash content than the other two scaling routines. Also observe that all six models were two-dimensional according to cross-validation.

Table 13.2: Summary statistics of six PLS models.

Model	R^2X	R^2Y	Q^2_{int}	Q^2_{ext}
Even, UV-scaling	0,934	0,849	0,828	0,794
Even, Par-scaling	0,977	0,842	0,824	0,796
Even, No scaling	0,983	0,837	0,820	0,797
Odd, UV-scaling	0,938	0,807	0,792	0,838
Odd, Par-scaling	0,980	0,805	0,793	0,832
Odd, No scaling	0,986	0,806	0,794	0,828

Although the overall performance statistics are numerically close, the models obtained are not similar. One way of understanding the properties of the various models is to plot PLS weights. Figures 13.3 – 13.8 show weights – for one component at a time – of the three models founded on the even-numbered training set (rows 1-3 in Table 13.2).

Obviously, the model established on the unscaled data has a first weight spectrum (Figure 13.5) strongly reminiscent of the raw data (cf. Figure 8.11). On the other hand, the PLS model founded on UV-scaled data produces a first weight spectrum bearing little relation to the measured spectra (Figure 13.3). It appears that UV-scaling up-weights the left- and right-hand parts of the spectral region, where incoming signals are weak and contain little information. We note that for spectral regions containing mostly noise, one would expect zero or near-zero PLS weights. This is not the case in Figure 13.3.

Interestingly, the Pareto procedure has produced a first weight spectrum (Figure 13.4) that is similar to the weight spectrum of the unscaled model (Figure 13.5). This indicates that Pareto scaling might be a sensible alternative for multivariate calibration based on spectroscopic data.

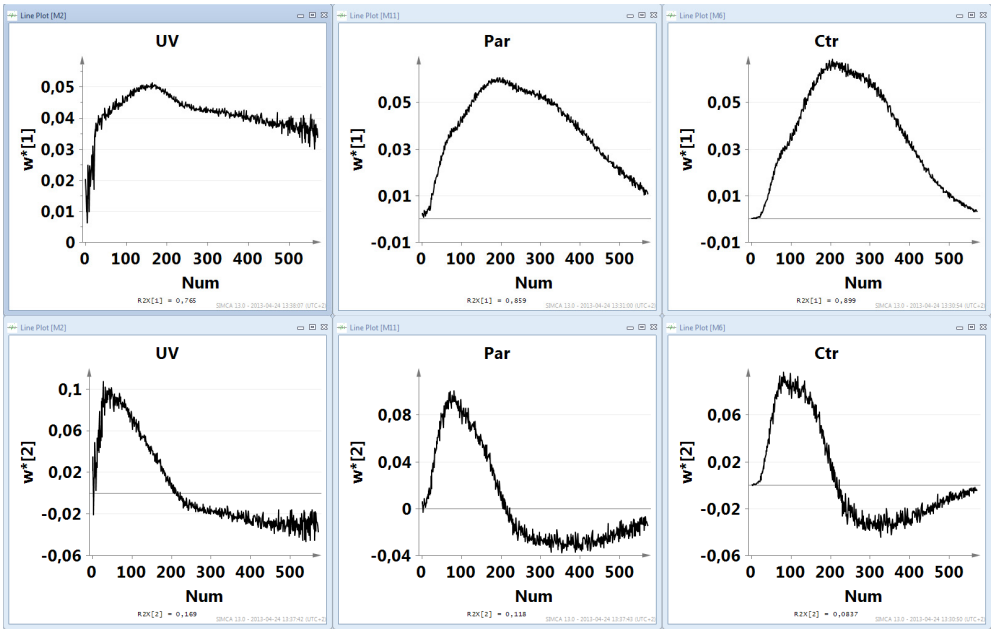


Figure 13.3: (upper left) PLS weight spectrum of first PLS component of sorted, UV-scaled sugar data. Figure 13.4: (upper middle) Same as in Figure 13.3, but with Pareto scaled data. Figure 13.5: (upper right) Same as in Figure 13.3, but with unscaled data. Figure 13.6: (lower left) PLS weight spectrum of second PLS component of sorted, UV-scaled sugar data. Figure 13.7: (lower middle) Same as in Figure 13.6, but with Pareto scaled data. Figure 13.8: (lower right) Same as in Figure 13.6, but with unscaled data.

Figures 13.6 – 13.8 reveal that the second components of each model are more similar to each other than are the corresponding first components. All three models describe a peak-like structure at low wavelengths. The somewhat noisy appearance of the second component weight spectra reflects the small amount of Y-variance explained by this component (10 – 12%).

In conclusion, in the current application, UV-scaling, Pareto scaling, or no scaling, produce calibration relationships of similar predictive ability. If predictive ability is of primary importance, any of the three scaling methods can be chosen. However, model transparency is better and model interpretation is easier with Pareto scaled or unscaled data.

13.5 Related scaling procedures

13.5.1 Block-scaling

One limitation of the scaling procedures discussed so far is that they do not consider whether variables are grouped in blocks of naturally related descriptors, or the number of variables in each such block. If, for example, UV-scaling is used, a large block of variables will dominate over a smaller block of variables, for purely numerical reasons. This is often not wanted.

We will now exemplify this problem. Consider a health related questionnaire completed by a number of people. In this questionnaire, there are ten questions regarding body fitness and exercising habits, 30 questions about food consumption patterns, five questions related to tobacco and alcohol usage, and three measured variables indicating body weight, blood pressure, and cholesterol level in blood.

The three measured variables probably convey more information about the health status of an individual than all the other variables put together. However, there is an apparent risk that the quantitative information of the measured variables will be masked by variation in the other 45 variables. With UV-scaling the former block would get a total variance of three, while the latter block is assigned a total variance of 45. Thus, even though scaling to unit variance has been performed, the questionnaire block will dominate over the measured one.

One way of addressing this problem is to employ *block-scaling*. Also referred to as *battery-scaling*, this corresponds to down-weighting blocks of variables in relation to a selected basis scaling procedure. The basis scaling method is generally UV-scaling, especially when variables are markedly different in nature and numerical range. However, in multivariate calibration, procedures like no scaling or Pareto scaling may well be used as the basis for block-scaling.

Block-scaling can be done in many ways. In our experience, it is convenient to distinguish between soft and hard block-scaling. In soft block-scaling, each block of variables is scaled such that the sum of the variable's variances (after completed scaling) equals the square root of the number of variables in that particular block. Here, the additional scaling weight used is $1/(k_{block})^{1/4}$ – where k_{block} represents the number of variables in a block – which is multiplied by the basis scaling weight. Hard block-scaling involves even further down-weighting. With this approach the variables in a block are scaled such that the sum of their variances is unity. Here, the additional scaling weight used is $1/(k_{block})^{1/2}$.

We will now exemplify the impact of block-scaling when UV-scaling is selected as basis scaling method. The example that we have chosen deals with multivariate characterization of the 20 coded amino acids, reported by Hellberg and coworkers [Hellberg, *et al.*, 1986; Hellberg, *et al.*, 1991]. In order to quantitatively account for the properties of the 20 amino acids, a battery of 29 physico-chemical descriptors was collected. Among the descriptors used we find molecular weight, acidity constants, ^{1}H and ^{13}C NMR shifts, partition coefficients, and nine HPLC retention indices. Hellberg and coworkers used PCA to summarize this multi-property matrix in terms of three quantitative scales (PCA scores), denoted z_1, z_2, and z_3. The derivation of these z-scales is reviewed in the supplementary book material available on www.umetrics.com.

The nine HPLC variables form a group of naturally related variables. In order to prevent them from dominating the derivation of the z-scales, Hellberg and coworkers applied block-scaling. For illustrative purposes, we used the amino acid data and computed three PCA models, the first one based on UV-scaling, the second on soft block-scaling, and the third on hard block-scaling. Scores and loadings pertaining to the first two model dimensions of these models are plotted in Figures 13.9 – 13.14.

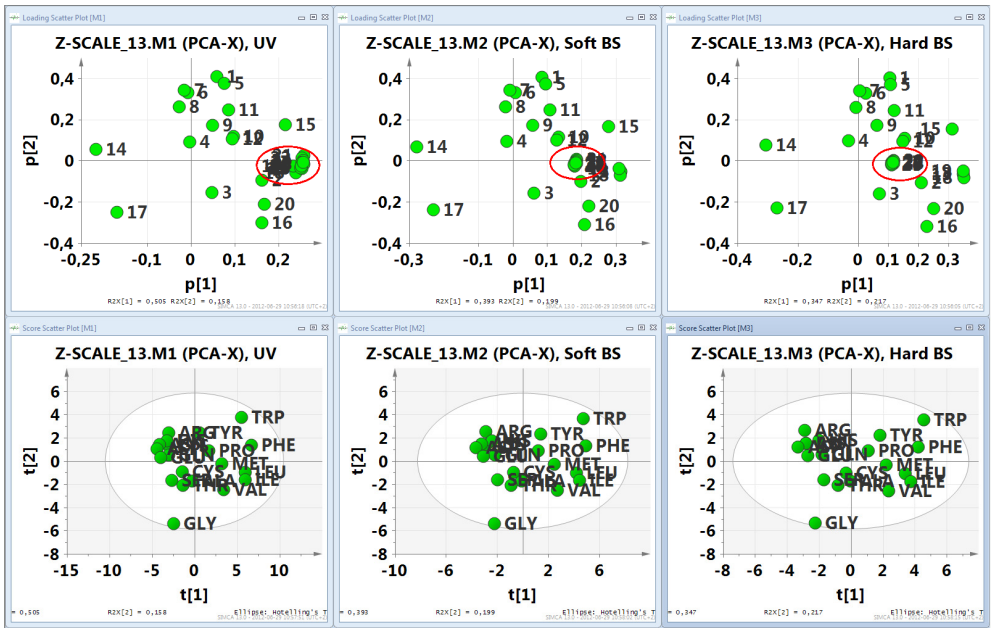


Figure 13.9: (upper left) PCA p_1/p_2 loading plot of model for amino acid data. UV-scaling was used. The location of the nine HPLC-variables is indicated by the circle. Figure 13.10: (upper middle) PCA p_1/p_2 loading plot of model for amino acid data when soft block-scaling was used. The nine HPLC-variables are marked. Figure 13.11: (upper right) PCA p_1/p_2 loading plot of model for amino acid data when hard block-scaling was utilized. The positions of the nine HPLC-descriptors are highlighted by the circle. Figure 13.12: (lower left) PCA t_1/t_2 score plot corresponding to Figure 13.9. Figure 13.13: (lower middle) PCA t_1/t_2 score plot corresponding to Figure 13.10. Figure 13.14: (lower right) PCA t_1/t_2 score plot corresponding to Figure 13.11.

The loadings following UV-scaling (Figure 13.9), soft block-scaling (Figure 13.10), and hard block-scaling (Figure 13.11) indicate the impact of each scaling technique. Obviously, with regard to the nine HPLC-variables, UV-scaling is weakest and hard block-scaling toughest. This behavior arises from the different variances after scaling of the HPLC-block. With UV-scaling the joint variance of these nine variables is nine, with soft block-scaling three, and with hard block-scaling unity.

The above results lead us to the insight that by selecting the scaling technique, it is possible to regulate the impact of a certain group of variables. However, regardless of which scaling approach is selected in the amino acid example, the resulting score plots remain surprisingly similar (cf. Figure 13.12 – 13.14). This indicates the stability of the existing multivariate characterization.

The use of the three z-scales in peptide QSAR is discussed in the supplementary book material available on www.umetrics.com.

13.5.2 Scaling, but no mean-centering

Occasionally, mean-centering should be avoided in data pre-treatment, because it might partially remove interesting information. Consider the PCA score plot displayed in Figure 13.15. This plot shows the relationship among 57 patients prior to and after a four-week treatment at a health-care center. With this plot it is possible to overview changes in health status and physiotherapeutic test results due to treatment. As seen, each individual occurs twice in the score plot, because the X-matrix has 114 rows. It is possible to connect data before (F) and after (E) treatment by drawing arrows. For clarity only a few arrows have been drawn. The length of each arrow indicates the impact of treatment on each patient.

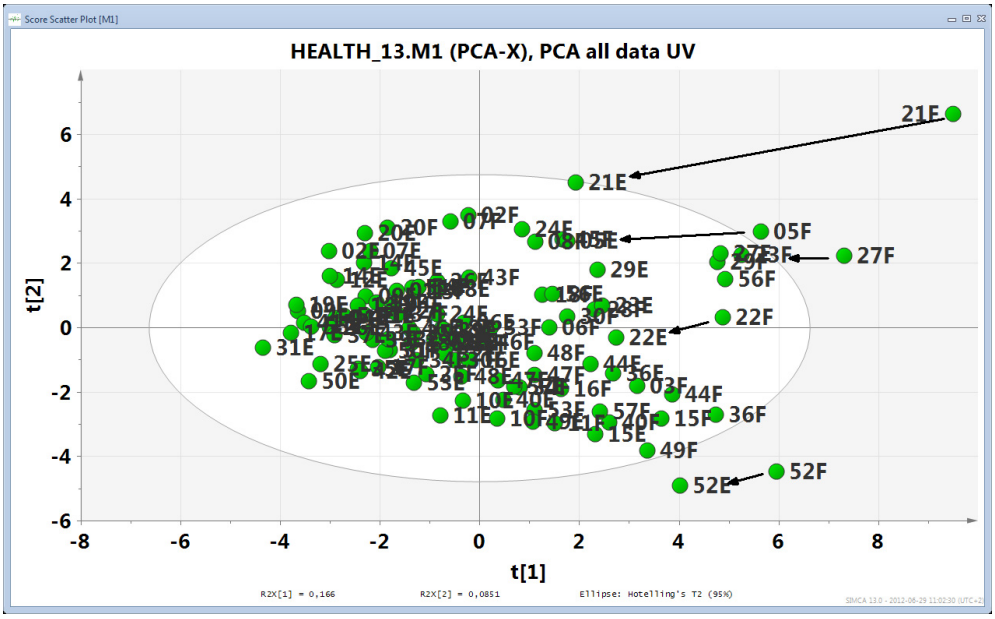


Figure 13.15: PCA of UV-scaled and mean-centered health-care data. Because the observations are paired, each patient will appear twice in the score plot, according to health status and physiotherapeutic test values before (F) and after (E) treatment. It is possible to draw one arrow per subject, representing the change due to the treatment undergone. For clarity, only a few arrows have been inserted in the score plot.

Because these observations are paired, an alternative way of analyzing the data would be to form new *difference* variables reflecting the numerical changes in the original variables before and after treatment. A PCA model can then be computed based on UV-scaled but uncentered data. In this case, data should *not* be mean-centered, as this would partly eliminate the interesting effect, the magnitude of change in a variable due to treatment.

Figure 13.16 shows a score plot of the first two components obtained when applying PCA to a matrix of 40 difference variables. To improve clarity, Figure 13.17 is a magnification of parts of Figure 13.16. If there were no treatment effects, all observations would appear as a cluster around the origin of the score plot. Here, however, we find that all subjects have

moved to the left along the t_1-axis. Evidently, patient 21 was most susceptible to the treatment given.

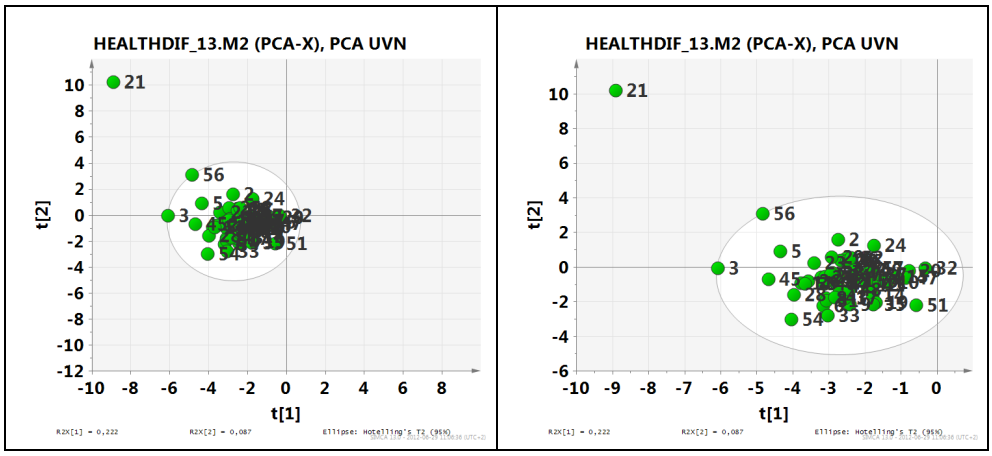


Figure 13.16: (left) PCA score plot of matrix of difference variables. Here, data were scaled to unit variance, but not corrected for the mean. The response to treatment of each individual is indicated by the distance to the plot origin. Figure 13.17: (right) Magnification of preceding graph.

A contribution plot can be constructed to interpret which variables make observation 21 extreme in the t_1/t_2-plane. We chose to make a contribution plot between observation 21 and observation 32, as the latter lies close to the plot origin (Figure 13.18). As shown by Figure 13.18, the striking difference between observations 21 and 32 is variable TY, which reflects difficulty in breathing. Observation 21 has a larger difference in variable TY than observation 32. This means that the primary determinant for the remote positioning of subject 21 is that his/her breathing properties have improved substantially.

A general model interpretation can be accomplished by considering the loadings plotted in Figure 13.19. In so doing, it should be remembered that all subjects, except no 21, are scattered mainly along the first component. The model interpretation suggests that variables related to cholesterol level (CO) and body-mass-index (BM) have decreased as a result of the stay at the health-care center. In addition, the loadings reveal that increases have occurred in physical fitness (TV) and high-density blood lipids (HD).

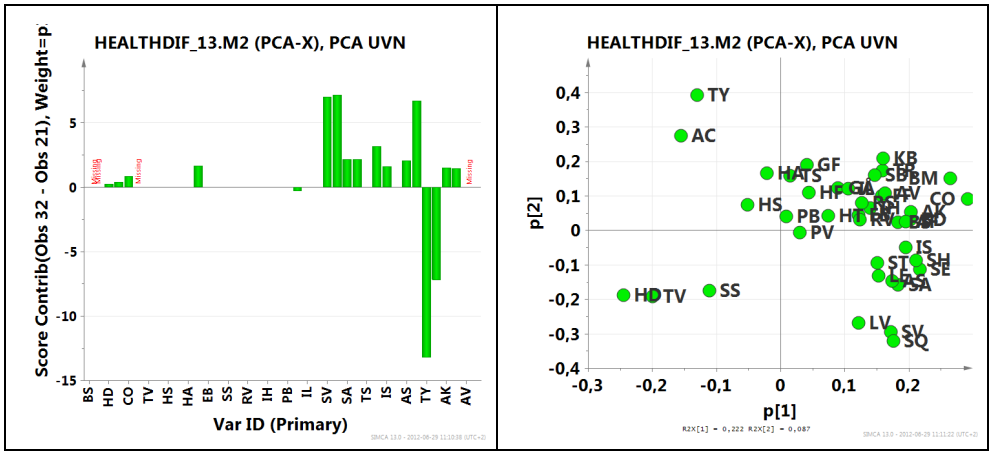


Figure 13.18: (left) Contribution plot (scores mode) between observations 21 and 32. When going from observation 21 to 32, variables SV, SQ and LV increase, and variables TY and AC decrease. Figure 13.19: (right) PCA loading plot corresponding to Figure 13.16. Because the observations are mainly scattered to the left along the horizontal plot direction, variables with high loading in the first component contribute most strongly to the separation. Hence, variables BM (body-mass-index) and CO (level of cholesterol) have decreased, and variables HD (amount of high-density blood lipids) and TV (physical fitness, test value) have increased as a result of the treatment.

13.5.3 Double-centering

Another common method for data pre-treatment is *double-centering*, that is, centering both columns and rows of a matrix. A typical example where double-centering is often used is in the generation of a map from a distance matrix. Such a distance matrix could be a simple table expressing distances between different geographical locations (Figure 13.20), or a similarity matrix between compounds in multivariate space. For further reading, please see the excellent book of Jackson [Jackson, 1991].

	ATA LANTA	CHI CAGO	DEN VER	HOUS TON	LA	MIA MI	NEW YORK	SAN FRAN
ATLANTA	0	587	1212	701	1936	604	748	2139
CHICAGO	587	0	920	940	1745	1188	713	1858
DENVER	1212	920	0	879	831	1726	1631	949
HOUSTON	701	940	879	0	1374	968	1420	1645
LA	1936	1745	831	1374	0	2339	2451	347
MIAMI	604	1188	1726	968	2339	0	1092	2549
NEWYORK	748	713	1631	1420	2451	1092	0	2571
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571	0

Figure 13.20: A distance table from which to generate a map. Data analyzed by C(F)A often display this kind of lay-out.

Double-centering is the basic transformation of data that is needed for the method of *correspondence analysis,* CA. CA is similar to PCA with a specific post- and pre-treatment of the X-matrix. CA has found widespread use in applications dealing with species abundance data in eco-toxicological monitoring [Massart, 1997]. In pollution monitoring around, say, an offshore oil production platform, counts of many species are registered at measuring stations along a pollution gradient. Often such species abundance curves show a unimodal behavior along a concentration gradient, that is, the number of specimens often first rises and then falls. A hypothetical and somewhat extreme example is formulated in Figures 13.21 and 13.22.

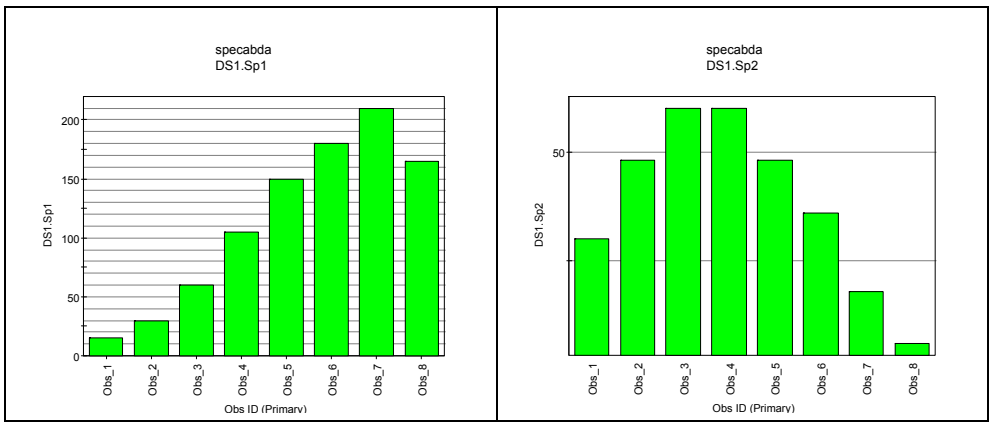


Figure 13.21: (left) A fictitious example showing a unimodal evolution of species 1 for eight measuring stations (observations) along a pollution gradient. Figure 13.22: (right) Graph similar to Figure 13.21, but for another species, species 2.

When analyzing many species abundance curves it is often found that they are not congruent, i.e., not synchronized, because of the different sensitivities different species have to the underlying environmental factors. As we shall see below, this might pose an interpretation problem if PCA is used for analysis.

By plotting the counts of species 1 and 2 against each other, Figure 13.23 is obtained. The extraction of the first PC now implies finding the line which best explains the data. The location of the first PC is indicated by the dotted line in Figure 13.23. Apparently, a one-component model does not explain all the variation in the data.

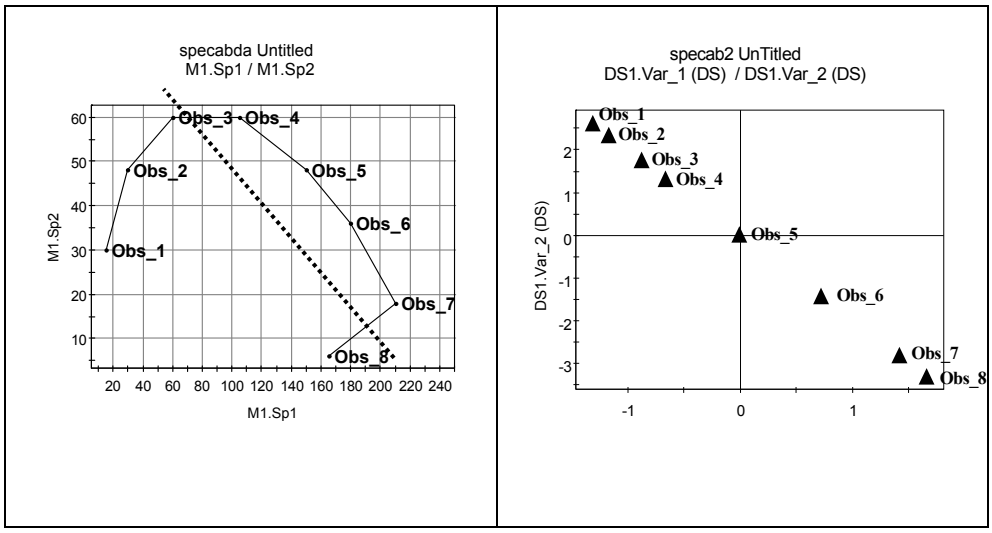


Figure 13.23: (left) Abundances of species 1 and 2 plotted against each other across the eight measuring stations. The solid line represents the order in which the measuring sites appear along the pollution gradient. The location of the first PC is indicated by the dashed line. By projecting the observations onto this line, the sequence 3, 2, 4, 1, 5, 6, 8, 7, is obtained. Figure 13.24: (right) Plot of "species counts" modified by correspondence analysis (hypothetical case). A straight-line model is now adequate for describing the data, and the problem of sequence inversion has disappeared.

However, what eco-toxicologists find more troublesome is that when projecting the observations onto the dashed line, an *inversion* of measuring sites occurs (Figure 13.23). If the first PC was mistakenly interpreted as representing a pollution gradient, we would infer that measuring station 3 is less polluted than stations 1 and 2, although the reverse is true. Such inversion may also occur at the other end of a pollution gradient.

In this context, correspondence analysis is applied with the aim of alleviating the problem of sequence inversion. This is done by a special pre-treatment of the X-data [Jackson, 1991]. Figure 13.24 shows a scatter plot indicating a typical outcome of CA, i.e., the modified species counts that result from CA better agree with the order among the measuring stations, and thereby these scales better conform with the pollution gradient. For further reading, reference is made to the work of Massart and Kvalheim [Massart, 1997].

It must be stressed that CA does not always solve this type of problem, also known as the *horseshoe* problem, as well as described in the fictitious example. The sketched example is somewhat extreme in that it only considers counts of two species. Often counts of many species are available, and often high percentages of missing data occur, which may make the data analysis more complex.

Finally, we wish to emphasize that in many applications, the removal of a curved data structure, like the one displayed in Figure 13.23, is to be discouraged. Such curved behavior is often part of the information residing in a data table, e.g., that operating conditions of a production plant vary over time.

13.6 Trimming and Winsorizing

The performance of most techniques used for pre-processing of data, i.e., methods for centering, scaling, transformation, filtering, etc., is critically related to the properties of the raw data. These methods work well only when data are homogeneous and devoid of influential outliers. Outliers in raw data are observations with extreme numerical values. The likelihood of encountering outliers increases as the number of data points increases. The presence of outliers in raw data will influence the estimates of location and dispersion parameters (average and standard deviation), i.e., the kind of parameters that are used in centering and scaling for example. Fortunately, there are simple pre-processing techniques that give fairly robust estimates of location and dispersion parameters, simply by protecting against the presence of gross outliers.

A simple and intuitively appealing procedure for robust estimation of averages and standard deviations is called trimming [Barnett and Lewis, 1994; Wold, *et al.*, 2002; Kettaneh, *et al.*, 2005]. A fraction α of lower and upper extreme values of a variable are discarded (and replaced with blanks, i.e. "missing data") before the analysis. Hence with trimming, we are not particularly concerned with the outliers, but only want to remove them in order to be able to model the systematic structure in the data better. Different variables may be trimmed using different levels of α, and the trimming may or may not be symmetrical (i.e., using the same α for both tails of the variable distribution). Additionally, instead of using a fixed level of α, the trimming can be carried out using lower and upper thresholds given by the standard deviation or a user specified limit.

An alternative to Trimming is a method called Winsorizing [Barnett and Lewis, 1994; Wold, *et al.*, 2002; Kettaneh, *et al.*, 2005]. Winsorizing is similar to trimming, except that instead of leaving empty entries in the data table (trimming), the blanks created when we remove the extreme data points are replaced by the values of the nearest observations or some other reasonable values. Thus, with Winsorizing we try to accommodate outliers, by diminishing their leverage in the forthcoming model. Winsorizing may use the same level of α as trimming does. It is also possible to Winsorize below and above some fixed lower-tail and upper-tail values (e.g., standard deviation or user entered limit). Below, Figures 13.25 - 13.27 show the effect of Winsorizing.

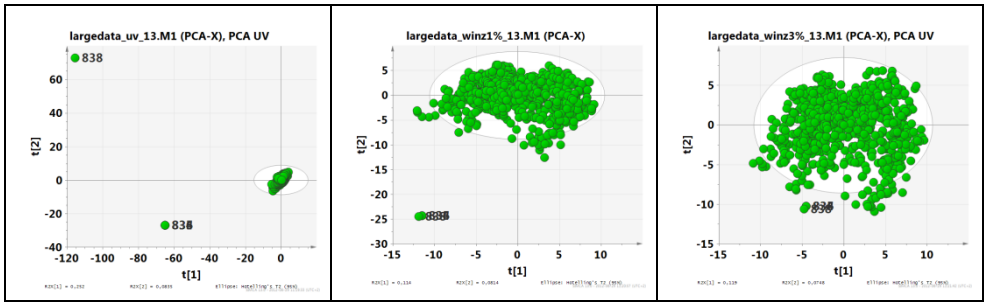


Figure 13.25: (left) PCA score plot of a process data set containing three strong outliers (two are overlapping). This data set has 834 observations and 160 variables. Figure 13.26: (middle) After Winsorizing the three outliers are much closer to the main cluster in the PCA score space. This enables a better understanding of the properties of the main cluster. The data set was Winsorized by removing the upper and lower 1% of the variable distributions and replacing those numbers by the values at the respective 1%-levels. Figure 13.27: (right) Same as in Figure 13.26, but with the 3%-levels used instead of the 1%-levels. The impact of the outliers has diminished.

The methods of Trimming and Winsorizing therefore affect the estimation of averages and standard deviations, as well as the subsequent multivariate models, all of which will be much less influenced by the presence of gross outliers. In order to interpret these outliers, we can trim the training set and thereafter bring back the untrimmed data as a prediction set. Then the model is not strongly influenced by the outliers, and they can, in original form, be scrutinized during the prediction phase.

13.7 Questions for Chapter 13

1. What is a scale?
2. What is scaling?
3. Why is scaling necessary?
4. When is scaling to unit variance appropriate?
5. How does Pareto scaling relate to UV-scaling and no scaling?
6. When is no scaling, but mean-centering of interest?
7. When is scaling, but no mean-centering of interest?
8. What is block-scaling?

9. When should block-scaling be applied?
10. What is Trimming? What is Winsorizing? When are these methods useful?

13.8 Summary and discussion

In Chapter 13, scaling of data has been considered. PCA, PLS, OPLS and O2PLS are sensitive to the choice of scaling method. This is both an asset and a limitation. Provided that scaling is applied in a cautious and careful manner, it can enhance the quality of data analysis. Prior knowledge about the relevance of a variable should be reflected in the scaling. In principle, however, if scaling is done carelessly, almost any result can be obtained. The relevance of a model founded on inappropriately scaled data is questionable.

14 Signal correction and compression

14.1 Objective

The objective of this chapter is to discuss methods for *signal correction* and *signal compression*. Signal correction and compression are of great interest in multivariate calibration, but may be used in many other fields, as well. Spectral data are often pre-processed ("corrected") in order to enhance the predictive power of a calibration model, because variation in X that is unrelated to Y may weaken the forecasting ability. Compression of spectral data is an area of rapidly growing interest, particularly in process applications, where large masses of data are usually handled. Wavelet analysis is the method discussed here for use in signal compression. Spectral filters may be combined, "chained", and one such possibility is hinted at using an example dataset consisting of FT-IR measurements on emission gases from a combustion process.

14.2 Introduction

One major objective in process monitoring is to determine and predict product quality. Product quality is typically a complicated property and often multivariate data are gathered to describe the quality of a given product in a given application. Spectroscopic techniques, such as near-infrared (NIR) and fluorescence measurements, are becoming increasingly used for this purpose. Usually, spectral data are used in conjunction with multivariate calibration for modeling the quality of process outputs.

In multivariate calibration, a quantitative relation is established between the digitized spectra (the X-data) and the properties (the Y-data) of the samples of interest. Examples of modeled properties might be thickness of a tablet coating, ash content in sugar production, strength of a polymer, ethanol content in wine, or viscosity of a solution. For extracting the predictively meaningful information from spectroscopic data, partial least squares projections to latent structures (PLS) is a widely used modeling technique. Lately, the OPLS method is increasingly used due to its interpretational advantages.

In order to enhance the predictive power of multivariate calibration models, spectral data are often pre-processed ("corrected") prior to data analysis. This is because variation in X that is unrelated to Y may degrade the predictive ability of a multivariate calibration model. Pre-processing of spectral data can be pursued along different trails, but the common denominator of most methods is the objective of removing undesired systematic variation in the data. Such undesirable systematic variation might be, for instance, baseline drift, multiplicative scatter effects, and wavelength regions of low information content.

Common approaches for pre-processing of spectral data are multiplicative signal correction (MSC) [Geladi, *et al.*, 1985], standard normal variate (SNV) correction [Barnes, *et al.*, 1989; Dhanoa, *et al.*, 1994], Savitzky-Golay smoothing [Savitzky and Golay, 1964], and first and second order derivation. A good discussion on the geometry of SNV and MSC was recently reported by Tom Fearn and coworkers [Fearn, *et al.*, 2009]. Moreover, the extended MSC, EMSC, is described and illustrated by Harald Martens and coworkers [Martens, *et al.*, 2003]. All these signal correction methods are interpretable as different cases of filtering. A filter is a mathematical function through which a signal, e.g., a fluorescence spectrum, is processed in order to "improve" its properties. What is meant by "improved properties" may

vary from application to application, and in many cases vague arguments are used to define the objective of signal filtering.

The expression “improved properties” is best illustrated by an example where the objective is signal de-noising. Consider Figures 14.1 and 14.2, which show the same “signal” before (Figure 14.1) and after (Figure 14.2) data pre-treatment. Apparently, the “signal” has obtained “improved properties” as a result of the data pre-processing. The underlying data are UV-spectra (diode array) from an HPLC application in the pharmaceutical industry (each row is a spectrum and each column is comparable to a traditional chromatogram). PCA was applied to summarize the data, and the “signal” displayed in Figures 14.1 and 14.2 is the second score vector t_2. The signal enhancement was accomplished by setting the scaling weight of a group of noisy variables to zero (“zeroing”). This approach is often called *information scaling* and is related to the procedures discussed in Chapter 13.

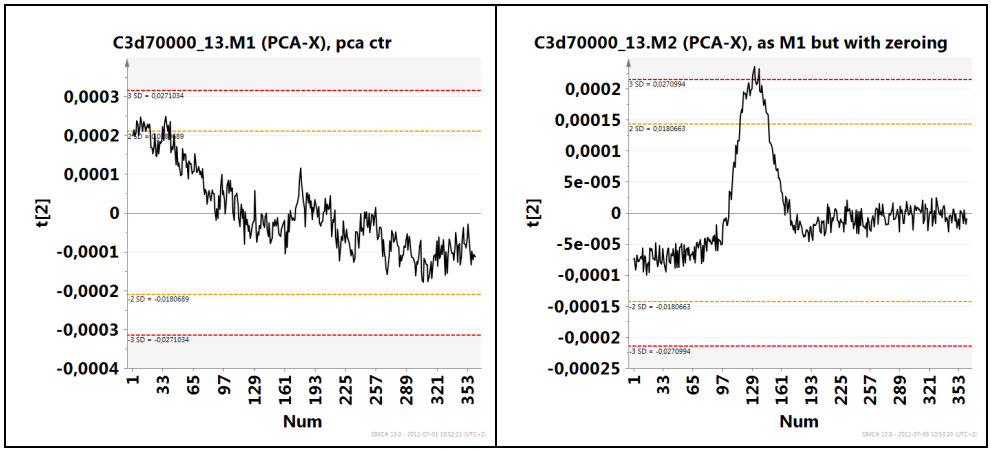


Figure 14.1: (left) PCA score vector t_2 of HPLC diode array data prior to zeroing. Figure 14.2: (right) PCA score vector t_2 after zeroing. A previously unseen chromatographic peak is now discernible. The second PC is less noisy after zeroing.

In order to more precisely define the objective of signal filtering, Wold and coworkers developed a filtering technique called orthogonal signal correction (OSC) [Wold, *et al.*, 1998a]. In this approach, the objective is mathematically well-defined, namely, that what is removed from the X-data should be mathematically independent of Y. OSC is a PLS based solution that removes from the X-data (the matrix of predictors) variation that is unrelated to Y (the matrix of responses, or a single response variable y). This is possible because OSC uses Y to construct a filter of X. The result of OSC is a model based on one or more PLS components conveying information about the correction of X. Conventional PLS diagnostics, such as, scores, loadings, and weights are then available to interpret exactly which kind of information has been removed from X. It should also be realized that the OPLS method, which is outlined in Chapter 5, includes the filtering capability of OSC as an integral part of the model.

Furthermore, in the context of multivariate calibration, wavelet analysis is gaining more and more attention. Alsberg and coworkers have provided a good introduction to wavelet analysis and have also highlighted a number of chemical applications where wavelet analysis could be of interest, e.g., for de-noising of IR spectra, for feature extraction in the classification of NIR spectra, and for noise suppression and data compression of NIR data [Alsberg, *et al.*, 1997]. Additionally, Trygg and Wold reported on the use of wavelets and PLS for multivariate calibration of compressed NIR spectra [Trygg and Wold, 1998]. A wavelet based multiscale extension of the PLS algorithm for advanced online process monitoring was reported by Lee and co-authors [Lee, *et al.*, 2009].

14.3 Methods for signal correction and compression

In this section, a brief explanation is given of four methods used for signal correction. These methods are orthogonal signal correction (OSC), multiplicative signal correction (MSC),

standard normal variate (SNV) correction, and first, second and third order derivation. Figures 14.3 – 14.8 show the impact of some of these filters on the Carrageenan NIR dataset [Dyrby, *et al.*, 2004]. We also describe an approach to wavelet analysis, which is deployed for signal compression. Wavelets may also be used for signal correction purposes.

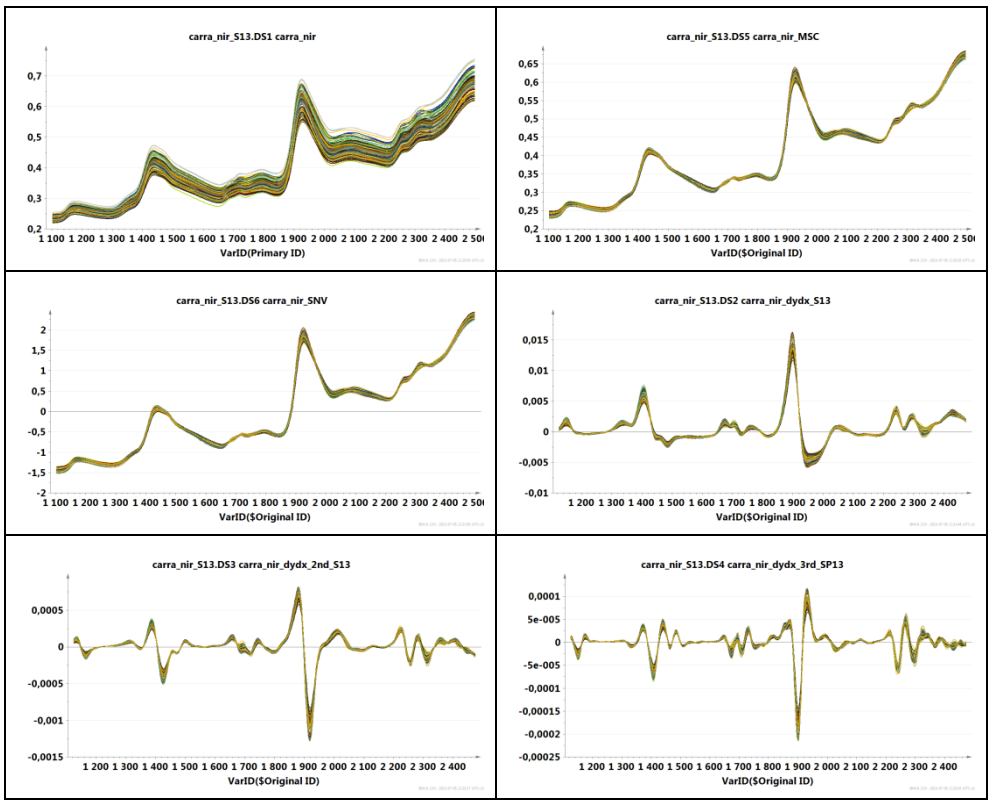


Figure 14.3: (upper left) NIR spectra (1100-2500 nm) of Carrageenan samples – Raw data. Figure 14.4: (upper right) Spectra after MSC filtering. Figure 14.5: (middle left) Spectra after SNV filtering. Figure 14.6: (middle right) Spectra after taking the first derivative. Figure 14.7: (lower left) Spectra after taking the second derivative. Figure 14.8: (lower right) Spectra after taking the third derivative.

14.3.1 Orthogonal signal correction, OSC

We will now give an introductory account of the OSC algorithm [Wold, *et al.*, 1998a]. OSC is usually used to remove one component at a time from **X** using the NIPALS algorithm. This has the advantage that the approach will cope also with moderate amounts of missing data, as do ordinary PCA and PLS. Prior to calculations, **X** and **Y** can be transformed, mean-centered, and scaled according to standard procedures.

The first step in the calculation of each OSC component is the formation of the first principal component score vector of **X**. This score vector, **t**, can be seen as a "precursor correction vector", which then is orthogonalized with respect to **Y** to give **t*,** the actual correction vector. Subsequently, PLS weights (**w**) are computed such that **Xw** becomes as close as possible to **t***. The correction vector **t*** is then processed in the NIPALS algorithm through **X** to give an updated score vector **t**, which is again orthogonalized with respect to **Y**. This is iterated until convergence. This means that **t*** is directed towards the longest vector that is orthogonal to **Y** and which still provides a good modeling and prediction of **X**. After convergence, a loading vector **p** is computed and then **t*p'** is subtracted from **X**. A second component may then be calculated, and so on. The resulting residual matrix **E** constitutes the filtered X-matrix.

It should be observed that the OSC filter is prone to overfit and it should therefore be used with caution to avoid overly optimistic results in multivariate calibration studies. Another fact that might complicate the use and evaluation of OSC is that the acronym OSC is used by several chemometric programs in which the implementation of the method is not

necessarily the same. As mentioned in Section 14.2, OSC is nowadays superseded by OPLS.

14.3.2 Multiplicative signal correction, MSC

With multiplicative signal correction, each digitized spectrum ($\mathbf{x_i}$', row-vector in **X**) is regressed against the average spectrum (**m**), according to $x_{ik} = a_i + b_i m_k + e_{ik}$. From each spectrum one subtracts the intercept (a_i) and divides by the slope (b_i) to get the corrected data, according to $x_{i,corr}' = (x_i' - a_i) / b_i$. The result of MSC is that each "corrected" spectrum has the same offset and amplitude (Figure 14.4). With this formulation, we should realize that there is a risk of obtaining vectors **a** and **b** that are correlated to **Y**. Therefore, MSC may remove from **X** information that is relevant for modeling and predicting **Y**.

Martens and coworkers have published extensions and modifications of the basic MSC, denoted extended multiplicative signal correction (EMSC) and extended inverted signal correction (EISC) [Kjaer Pedersen, *et al.*, 2002; Martens, *et al.*, 2003]. The primary objective of these filters is to facilitate a removal of spectral artifacts and interferences that are uncorrelated to the target analyte concentration, i.e., to allow a separation of physical light-scattering effects from chemical and vibrational light absorbance effects in spectra.

14.3.3 Standard normal variate (SNV) correction

Both OSC and MSC, and their extensions, are model-based signal correction methods and use all spectra of the training set (calibration set) to derive the parameters of the signal filter. As such, these filters are sensitive to outliers, sub-groupings, non-linearities and other anomalies in the data. However, with homogeneous and representative calibration data these methods represent stable and useful filters.

Filters based on individual observations (spectra) are less sensitive to peculiarities in the raw data. Examples of such filters are standard normal variate correction (SNV), the first derivative spectrum, the second derivative spectrum and the third derivative spectrum.

The mathematical formulation of SNV is similar to that of MSC. In SNV, the parameters a_i and b_i are calculated as the average and the standard deviation of the i^{th} row of **X**. Actually, this corresponds to row-centering and scaling (compare column-centering and scaling). Figure 14.5 shows the effect of applying SNV to the Carrageenan NIR data. The filtered spectra are typically very similar to the MSC treated spectra, as the two transformations are closely mathematically related [Dhanoa, *et al.*, 1994; Argyri, *et al.*, 2013].

14.3.4 First, second and third order derivation

A rapid and often utilized method for reducing scatter effects for continuous spectra consists of using derivatives [Naes, *et al.*, 2002]. The first derivative spectrum is the slope at each point of the original spectrum. It peaks where the original spectrum has maximum slope and it crosses zero where the original has peaks. The second derivative spectrum is a measure of the curvature at each point in the original spectrum. Usually, this derivative spectrum is more similar to the original spectrum and has peaks approximately as the original spectrum, albeit with an inverse configuration [Naes, *et al.*, 2002]. The effect of the first derivative is usually to remove an additive baseline ("offset"), whereas the effect of the second derivative also involves removal of a linear baseline. The SIMCA software also supports taking the third derivative spectrum, which is a measure of the "wiggliness" at each point of the original spectrum. Figures 14.6 – 14.8 show the impact of first, second and third order derivation of the Carrageenan NIR data.

A problem with the above approach is that differencing may reduce the signal and increase the noise, thus producing very noisy derivative spectra. Realizing this risk, Savitzky and Golay (SG) proposed an improvement based on a smoothing approach [Savitzky and Golay, 1964]. SG-derivatives are based on fitting a low degree polynomial function (usually of quadratic or cubic degree) piece-wise to the data, followed by calculating the first, second or third derivative from the resulting polynomial at points of interest. SIMCA supports the SG-derivative option.

14.3.5 Chained filters

Spectral filters may also be combined in a sequential order ("chained") to promote the effect of one another. SIMCA allows chaining up to four filters. Popular combinations include derivatives and MSC, and derivatives and SNV.

14.3.6 Wavelet analysis

We will now provide a brief introduction to how we use wavelet analysis. Wavelets look like small oscillating waves, and they have the capability of probing a signal according to scale, that is, bandpass of frequencies [Alsberg, *et al.*, 1997; Trygg and Wold, 1998]. The characteristic features of this approach are good compression and de-noising of complicated signals. The wavelet transform uses a mother wavelet, that is, a basis function, with a certain scale (width of the analyzing function window) to investigate the time-scale properties of an incoming signal. By varying the width of this window, both sharp and coarse properties of the signal are captured. A narrow wavelet is used for detecting the sharp features, and a wider wavelet is useful for uncovering general signal properties.

The mother wavelet can be selected from many different families of filters. Some examples are shown in Figures 14.9 – 14.17. The shape of the wavelet filter depends on the selected family and the order.

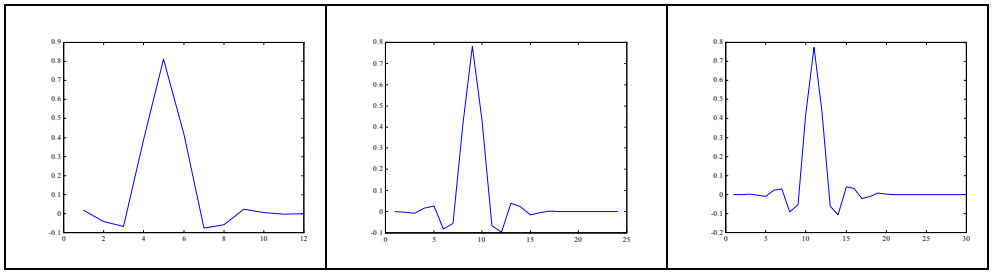


Figure 14.9: (left) Coiflet, order 2. Figure 14.10: (middle) Coiflet, order 4. Figure 14.11: (right) Coiflet, order 5.

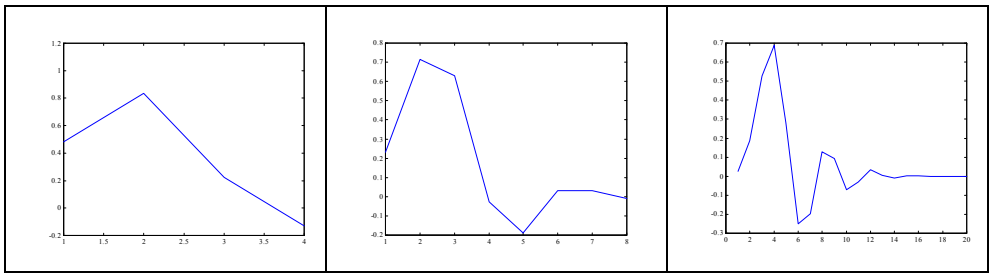


Figure 14.12: (left) Daubechies, order 4. Figure 14.13: (middle) Daubechies, order 8. Figure 14.14: (right) Daubechies, order 20.

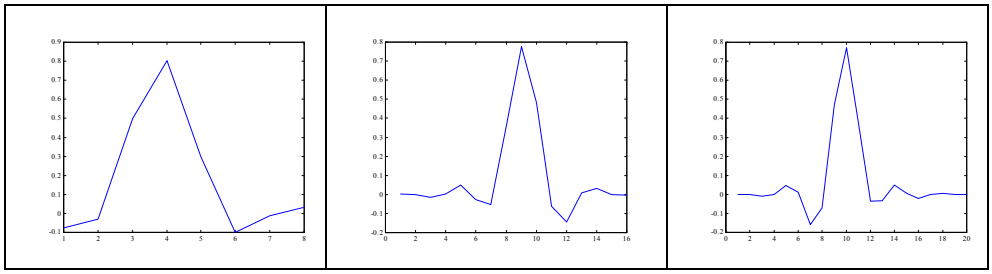


Figure 14.15: (left) Symmlet, order 4. Figure 14.16: (middle) Symmlet, order 8. Figure 14.17: (right) Symmlet, order 10.

The examples shown in Figures 14.9 – 14.17 all correspond to different families of orthogonal wavelet bases and wavelet order. Dau4 and Sym8 are two of the most widely used wavelets, and have appropriate properties for compressing smooth signals, like NIR and fluorescence data [Trygg and Wold, 1998].

In the compression of spectral data, two wavelet transform techniques emerge as very popular, i.e., the wavelet packet transform (WPT) with a best basis (BB) selection criterion, and the discrete wavelet transform (DWT) linked to the multiresolution analysis (MRA) approach (see description in [Trygg and Wold, 1998]). The DWT approach can be interpreted as a special case of the WPT approach. The former method uses a fixed orthogonal basis and is useful when the high-frequency part of a signal is uninteresting, as is often the case in the analysis of spectral data. The latter approach is useful when a high-frequency component is suspected, like in very peaked time series data.

14.3.7 Main steps in data compression

Figure 14.18 is a schematic showing the main steps involved in the signal compression (using DWT) as implemented in SIMCA. In DWT/MRA, the variance spectrum of the coefficient matrix (training set) in the wavelet domain is used in the compression. The positions representing a selected number of the largest variance coefficients are identified, and the corresponding columns from the wavelet coefficient matrix are extracted and fed into a compressed data matrix. In WPT/BB the average spectrum of the raw signals is first used to compute the "best basis". Thereafter the raw signals are transformed using the identified best basis, and the resulting wavelet coefficient matrix is handled in the same way as in the DWT approach.

In both approaches, the original positions of the extracted coefficients are saved. This is used in future compression of additional spectra, and in the reconstruction of loadings, coefficients, scores, and so on. To reconstruct the original loadings, coefficients, etc., the wavelet coefficients are placed in their original position in the wavelet vector. Subsequently, vacant positions are filled with zeros, and the inverse wavelet transformation is then conducted.

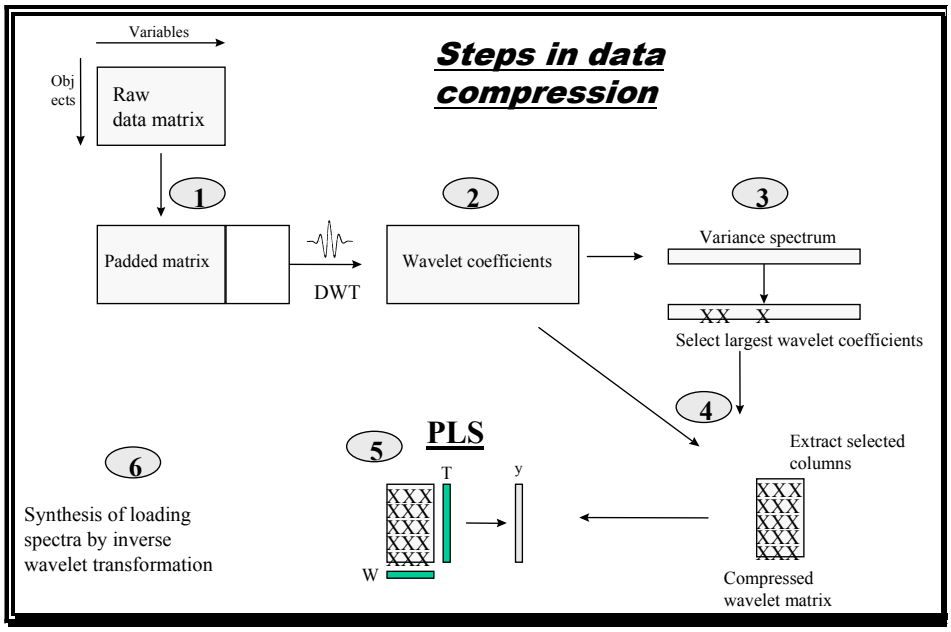


Figure 14.18: An overview of the steps involved in the compression of spectral data. (1) The raw spectra are padded to the nearest length 2^N, e.g. $2^{12} = 4096$ for a dataset comprising 4000 variables. (2) Each spectrum is transformed using a selected compression method, e.g., DWT or WPT, and a selected wavelet function of a given family and order, e.g., Dau4 or Sym8. (3) The wavelet variance spectrum is computed and the columns of the largest coefficients are identified. (4) The identified columns are extracted and stored in a compressed matrix. (5) PLS regression is performed based on the compressed X-matrix. (6) PLS scores and weights are reconstructed using the inverse wavelet transformation.

14.4 An example on chained filters

Some common spectral filters were introduced in the previous section and their effect on the Carrageenan NIR dataset was visualized through the series of plots in Figures 14.3 – 14.8. In this section, the objective is to provide a brief example on how two filters can be used in a sequence, i.e., chained. Filters can be chained for two main objectives. The first reason is that filters are combined to enhance one another in a synergistic fashion. Popular combinations involve MSC or SNV in combination with some derivative function. The second reason for chaining filters is to accomplish good compression of spectral data using wavelets. It is the latter aspect which is being dealt with here.

The example dataset consists of FT-IR measurements on outlet gaseous species arising in a combustion process. The total number of observations (time points, FT-IR spectra) is 41872 and the total number of variables is 3076 (wavenumbers in the region 400-5000 cm^{-1}). Altogether these measurements correspond to a sampling campaign lasting nearly 72h. Figure 14.19 shows the score control chart of the first principal component score vector for the model based on un-filtered spectral data. Evidently, there is a start-up phase with comparatively large changes. This initial phase lasts for about 1000 time points (spectra). Beyond the first 1000 time points (spectra), the process stabilizes somewhat, but there are still considerable jumps and periodicities with time. One apparent spike (outlier) is marked by a circle in Figure 14.19.

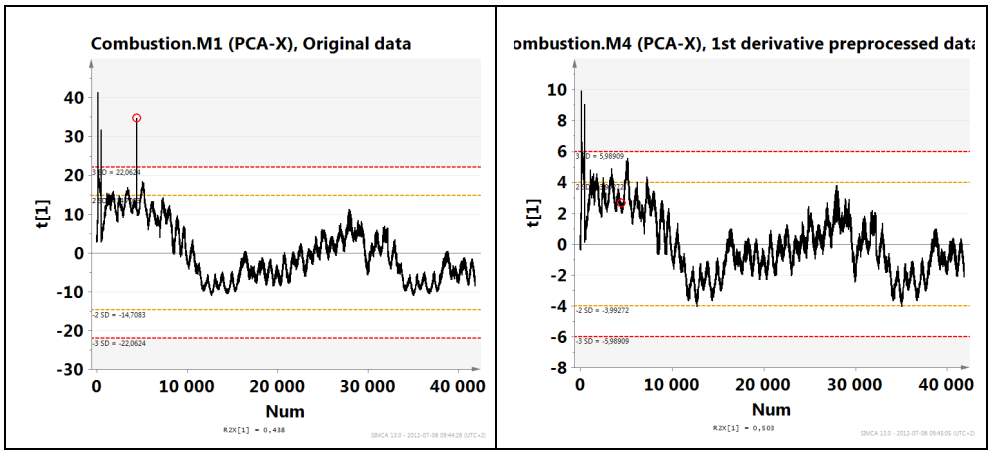


Figure 14.19: (left) Control chart of the first PCA score vector t_1 arising from the analysis of un-filtered spectral data. Pareto scaling of the data was used. The time span is approximately 72 hours. Figure 14.20: (right) Control chart of the first PCA score vector t_1 arising from the analysis of first derivative filtered spectral data. Pareto scaling of the data was used. The outlier noted in Figure 14.19 is now removed and the start-up variation is less pronounced in relation to the process variability in the remaining part of the sampling period.

Because of the large initial process variability and the presence of an outlier, it was deemed appropriate to work with spectral filters that are not sensitive to extreme samples (observations). Thus, the impact of the first derivative and the SNV filters was explored, the results of which are seen in Figures 14.20 and 14.21. The first score control chart from the PCAs on the filtered data show that (i) the outlier has been removed and that (ii) the initial start-up variability is somewhat suppressed, which effectively means we obtain a better resolution of the rest of the process variability. Thus, the filters used (first derivative and SNV) have accomplished a useful alignment of the original spectra, which can be corroborated by plotting the spectral data before and after filtering (no such plots shown).

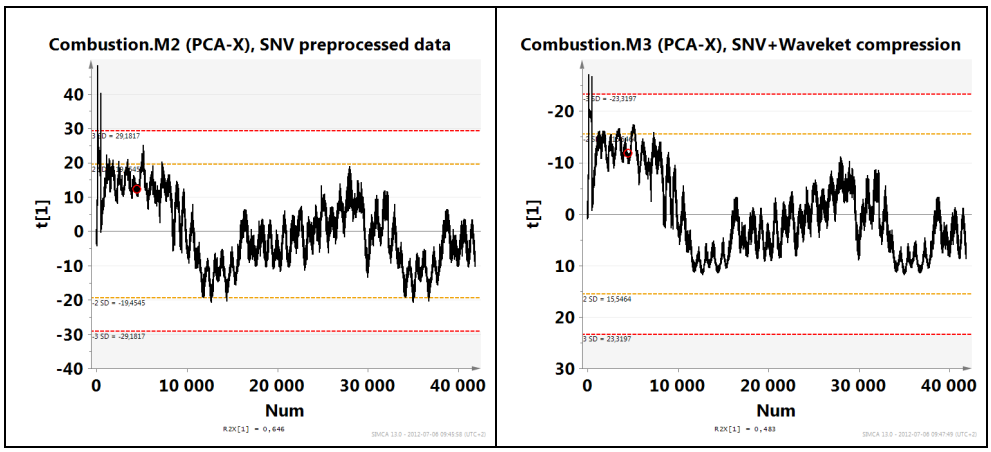


Figure 14.21: (left) Control chart of the first PCA score vector t_1 arising from the analysis of SNV filtered spectral data. Pareto scaling of the data was used. Figure 14.22: (right) Control chart of the first PCA score vector t_1 arising from the analysis of filtered and compressed spectral data. Pareto scaling of the data was used.

In the next step, wavelet compression was applied to the SNV filtered data. The Dau-4 wavelet was selected and by adhering to the steps schematically depicted in Figure 14.18, we extracted a total of 297 wavelet coefficients. These 297 wavelet coefficients retain 99.5% of the variance of the SNV filtered spectral data. Figure 14.22 shows the first score control chart of the PCA based on the filtered and compressed spectral data. Since the 297 wavelet coefficients account for 99.5% of the input variance, there is no loss of information following the data compression. Thus, it is possible to compress the dataset to less than 10% of its original size (297/3076 variables) using chained filters.

In parallel to the results just related, an alternative chained filtering was tested in which wavelet analysis was used to compress SNV filtered data even further. In this alternative procedure, as few as 142 wavelet coefficients were extracted, corresponding to 95% explained variance. From the appearance of the resulting PCA score control chart (no plot shown) it was concluded that not much information on the general behavior of the combustion process was lost. Hence, if desired, it is fully conceivable to compress the spectral measurements to less than 5% of their original size (142/3076 variables) without loss of generality and interpretability.

14.5 Questions for Chapter 14

1. What is filtering?
2. Why is filtering (signal correction) often necessary?
3. What is MSC? What is SNV? What are the results of these transformations?
4. What is OSC? What is the mathematical objective of OSC?
5. What is information scaling? When is it useful?
6. What is wavelet analysis? For which purposes can it be used?
7. How is it possible to evaluate the impact of signal correction and compression?

14.6 Summary and discussion

In this chapter, methods for signal correction and signal compression have been addressed. We have visualized the impact of MSC, SNV and derivatives on an NIR dataset. In this context, it is appropriate to point out that MSC can be implemented in different manners. In our approach we have used the whole spectral region for establishing the appropriate MSC parameters (slope and intercept of regression line between average spectrum and individual spectrum). Some MSC proponents have advocated that it should be applied to smaller and well-selected parts of the spectral region. However, it is often difficult to discover exactly

which parts of the spectral region will produce the best signal correction performance. Hence, we decided to use whole spectra.

Furthermore, the addition of wavelet analysis into the signal pre-processing scheme offered a substantial compression of the spectral data in the combustion process example. To some people, the wavelet approach is hard to understand and is associated with elaborate mathematics. In our opinion, the important thing to realize is that wavelet analysis produces a *transform* of the original data, from one numerical domain to another (see illustration in Figure 14.23). In this alternative domain, some signal manipulation is carried out, which may result in de-noising and compression. Subsequently, a reconstruction into the original signal domain is possible by reversing the steps of the wavelet approach.



Figure 14.23: A simplified picture of wavelet analysis (WA). Wavelet analysis is a transformation of the data from one domain to another. In the new domain, some manipulation is made to the signal, and then a reconstruction can be made into the original domain. The result is often a signal, or a set of signals, with improved properties. Using the Arabic numbering system, the number 21 is a little wiggled and not sharply printed. After transformation to the corresponding Roman numeral, some signal de-noising and sharpening have occurred. By removing the undesired "signal features" and reversing the steps of the transformation, it is possible to reconstruct the numerical value in the original domain (the Arabic numbering system).

Moreover, regarding wavelet analysis, a 2-D wavelet analysis approach for compression of spectral process data was proposed [Trygg, *et al.*, 2001]. The example concerned an industrial process monitoring situation, where near infrared spectra were measured on-line on top of a conveyer belt filled with wood chips at a Swedish pulp plant. This paper introduces a new parameter for outlier detection, the distance to the model in wavelet space (DModW). This parameter is analogous to the residual parameter DModX used in PCA and PLS. Additionally, the wavelet power spectrum (WPS) is introduced; it produces an instant overview of the time-frequency content in a signal. In the example given, wavelets improved the detection of spectral shift and compressed data 1000-fold without degrading the quality of the 2-D wavelet compressed PCA model.

Similarly to Trygg *et al.*, Teppola and Minkkinen also advocate the use of wavelets for multivariate analysis of complex industrial data [Teppola and Minkkinen, 2000; Teppola and Minkkinen, 2001]. Wavelets may also be used in multiscale analysis of process data [Bakshi, 1999; Lee, *et al.*, 2009].

15 MSPC

15.1 Objective

In Chapter 15 the objective is to discuss statistical process control, SPC. SPC aims to investigate whether a process is in a state of statistical control. A state of statistical control exists when key process variables or product attributes remain close to their target values, and do not change perceptibly. This kind of assessment is accomplished through use of control charts. An introduction to the most common control charts is given. SPC has a simple, yet natural multivariate extension, called multivariate statistical process control, MSPC. MSPC will be illustrated using the PROC1A and CUPRUM datasets.

15.2 Introduction

In Chapter 9 it was shown that multivariate data analysis is important in the monitoring of process behavior. Nowadays, modern process computers and electronic sensors are able to record massive amounts of process data, which are usually stored in huge databases. Such highly multivariate data require good analytical techniques like PCA, PLS, OPLS and O2PLS in order to separate systematic information from unsystematic variation. PCA, PLS, OPLS and O2PLS are relevant for process analysis because they are able to overcome the intrinsic problems of process data, problems related to (i) dimensionality, (ii) collinearity, (iii) noise and (iv) missing data [MacGregor, 1996].

One major objective of multivariate process modeling is to understand the current state of the process. Is the process operating as it should, or is there a tendency to depart from normal behavior? Let us consider an application introduced in Section 9.5, the PROC1A dataset. We remember that these data represent a chemical production plant, operated as a continuous steady-state process. There are 92 observations and 33 variables in the dataset. An unusual behavior started around time point 70; the process went out-of-control at time point 80 and had to be shut down.

The inspection of raw signals in terms of 33 time series plots is not very fruitful, especially when these are superimposed on top of one another in one single graph (Figure 15.1). Not even an experienced process operator with a trained eye can easily detect the process upset. Firstly, the overlay of many individual signals creates a graph that is difficult to overview and comprehend. Secondly, and more seriously, the information in the data is not appropriately exploited. The information resides in the combination of many signals, which can more clearly be expressed by a set of latent variables.

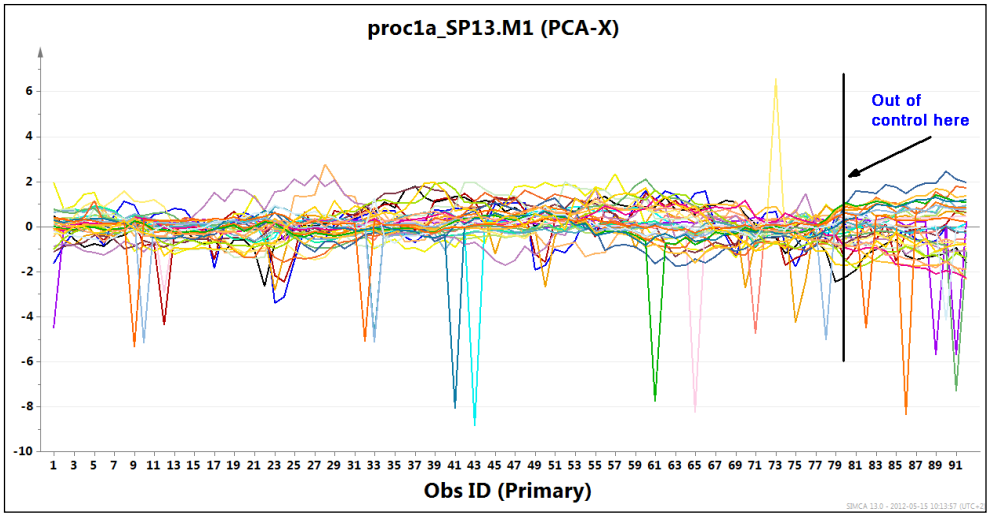


Figure 15.1: Plot of raw data of the PROC1A dataset. There are 33 trend curves superimposed on top of one another. Plotting the raw data does not reveal that anything strange is going on after time point 80.

Figure 15.2 shows the first two principal components of the PROC1A dataset. The control limit (the Hotelling's T^2 ellipse) was calculated using the information in all 33 variables. It is immediately seen that the process has migrated outside the tolerance ellipse after time point 80. Thus, this score plot is a very efficient way of accomplishing a process status visualization (recall that this objective of multivariate process modeling was discussed in Section 9.3.1).

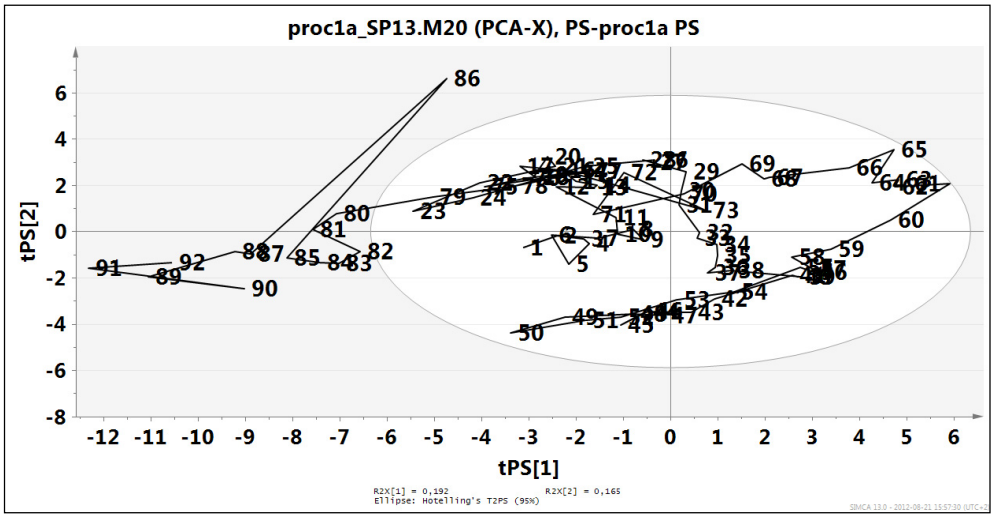


Figure 15.2: PCA t_1/t_2 score plot of the model of the PROC1A dataset. In the beginning, when the process behaves well, the observations wander around in a small region of the score plot. At the end, when the process behaves badly, a strong trend is expressed in the data.

Furthermore, the analysis of process data should not confine itself simply to an overview of the X-data by means of PCA. It is also relevant to explore relationships between the output variables, Y, and the input variables, X, using for example PLS, OPLS or O2PLS. The aim is to understand how to adjust X to maintain a target profile in Y, such as, an optimal combination of product quality properties. We note that product quality is typically a multivariate property and must be treated as such. Hence, a multitude of quality measurements should be made on the product, and these should be analyzed multivariately. The multivariate analysis will model the inherent complexity of the product quality in terms of a limited number of latent variables.

Latent variables derived from PCA, PLS, etc., are computations of process data that can be displayed graphically by means of control charts [MacGregor and Kourti, 1995]. For instance, an efficient way of displaying modeling results consists of plotting - in a combined

control chart - the extracted latent variables (scores) together with the residuals [Wikström, *et al.*, 1998a]. This enables the specification of multivariate (quality) control limits, rather than control limits connected to univariate control charts of the individual (quality) measurements. The use of such multivariate (quality) control charts will help ensure that the product is manufactured with good properties, according to all the monitored (quality) characteristics. This is the essence of multivariate statistical process control (MSPC).

The aim of this chapter is to introduce the MSPC concept and some associated control charting techniques. In doing so, we shall first get acquainted with the predecessor of MSPC, namely statistical process control, SPC. We will describe the objective and assumptions of SPC, and some basic types of control charts. Finally, we consider the natural and simple extension of SPC called MSPC. MSPC will be exemplified using the PROC1A and CUPRUM datasets.

15.3 Statistical process control (SPC)

15.3.1 Basic assumption of SPC

The basic principles of statistical process control were outlined by Walter A. Shewhart in the 1930s [Shewhart, 1931]. Over the years, SPC concepts and methods have proliferated and become very important in most industrial sectors. The objective of these methods is to monitor the performance of a process over time in order to understand whether the process behaves as it is expected to do. And if it does not, the analysis should provide clues to the problem, which enable process corrections to be made.

The basic assumption of SPC is that a process always adheres to a *state of statistical control*, unless a *special event* occurs. A state of statistical control exists when certain critical process variables and/or product attributes remain close to their target values, and do not change perceptibly. The only variation should be marginal everyday fluctuations that are essentially unavoidable and affect the process all the time.

15.3.2 Approach to detect the occurrence of a special event

In order to detect the occurrence of a special event, SPC charts, such as, Shewhart, CuSum (cumulative sum), and EWMA (exponentially weighted moving average) charts are routinely used to monitor key process variables and product attributes over time. This is accomplished by evaluating historical process data representing good past process behavior, and computing statistics like average and standard deviation (SD). Such information is then used for setting up control charts, e.g. by using the variable average as target and ± 3 SD.s as control limits, and referencing the current process situation against the historical one.

The principle of a control chart is shown in Figure 15.3. The numerical values of a variable are charted over time. It is usual to indicate the target value and its associated warning and control limits. In Figure 15.3 the variable average has been used to designate the target value, and the warning and control (action) limits are represented by 2 and 3 standard deviations, respectively. In SPC it is assumed that if the process is under control then the warning lines include 95% of all values and the control (action) lines approximately 99%. Discovering values outside the warning lines is interpreted as a warning that the process may be getting out of control. Finding values outside the control limits then indicates a need to bring the process back under control. As shown by Figure 15.3, the CUPRUM process is mostly under control according to the variation in the bismuth variable.

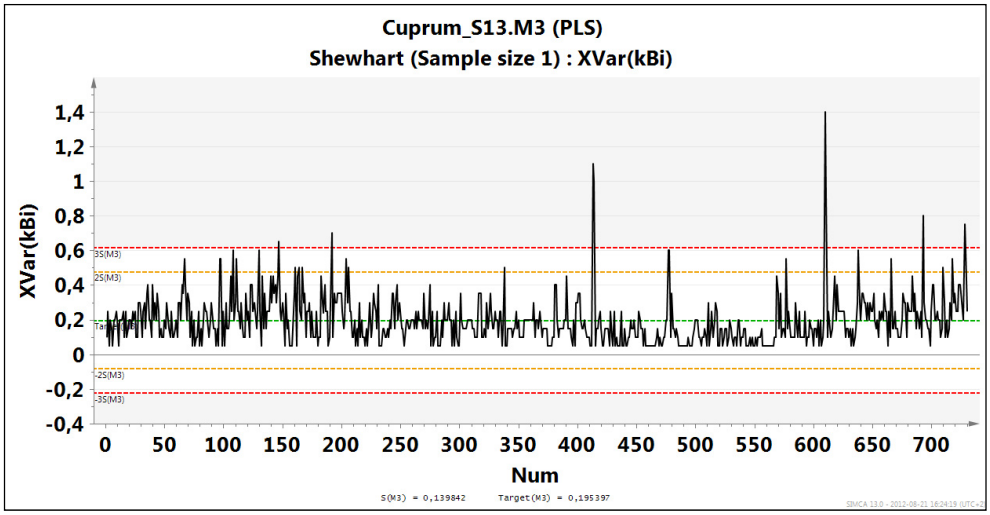


Figure 15.3: An example control chart. The plotted variable is amount of bismuth in copper (CUPRUM dataset). The x-axis is the observation number, which is equivalent to the time order. Two samples were taken each day during one year of copper production.

In SPC, control charts of individual key variables are often evaluated to find a *process upset* and an *assignable cause* for that special event. (In SPC-jargon finding an *assignable cause* means *interpreting deviations*). Once an assignable cause has been discovered, the idea is to undertake corrective measures, which will direct the process back to target. On a short-term basis this may lead to less rejected product, and in a long-term perspective significant process improvement might occur.

15.3.3 Scope of SPC

In Shewhart's day we were lucky to obtain one or a few measurements of product quality, and perhaps a few signals revealing the state of the process. Hence, it was natural to apply control charting methods to a limited number of variables and examine the resulting charts one by one. However, the problem with this approach is the inherently misleading nature of univariate control charts. Univariate control charts give the impression that process and quality variables are completely independent of each other, and therefore should be monitored in isolation. In reality, however, because many signals are measured on a process, the majority of which are often partially or strongly correlated, univariate control charts are inapplicable in many cases.

The above problem is best illustrated by an example. Consider Figure 15.4, where two quality variables have been plotted over time in separate control charts. For each variable, two measurements have been marked with a frame. According to the individual variables, these two samples are good in every respect since they are found well inside the control limits, located at ± 3 standard deviations. A completely different message is conveyed by the scatter plot of the two quality variables, which indicates that the two highlighted samples are outliers when the variables are considered together.

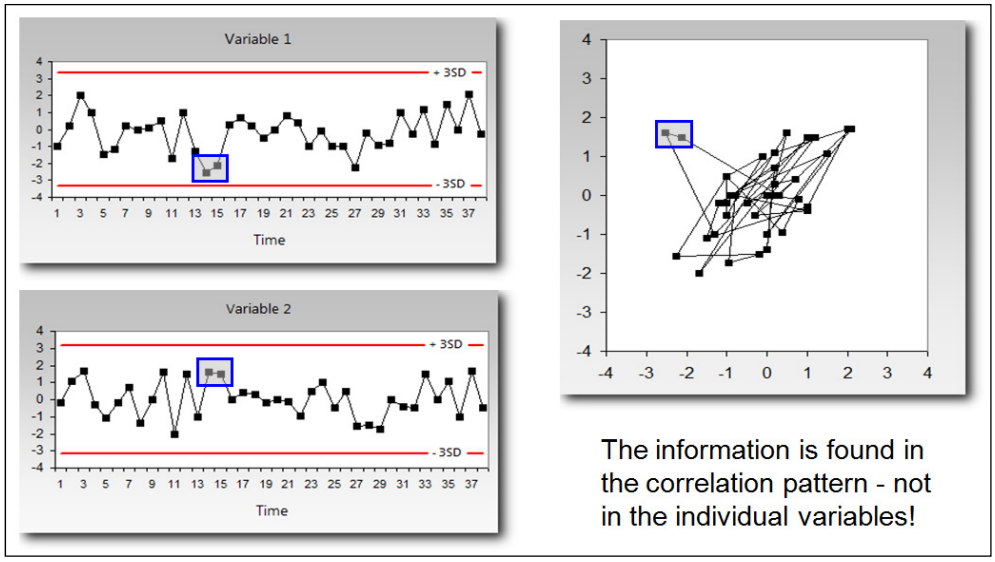


Figure 15.4: Two variables are charted over time, and for each variable two measurements have been marked with a frame. According to the individual variables, these two samples are inside the control limits, located at ± 3 standard deviations. In contrast, the two highlighted samples appear as outliers when the two variables are plotted against each other. Most outliers are NOT detected until all variables are analyzed together.

As shown by Figure 15.4, the information is found in the correlation pattern between the two variables, *not* in the individual signals themselves. In fact, most outliers are *not* detected until all variables are analyzed together. This is the key to early fault detection. As a consequence, one should be very careful when working with univariate control charts. It is very easy to be lulled into a false sense of security.

Conventional approaches to SPC using univariate control charts are usually not adequate for most modern process industries [MacGregor and Kourti, 1995]. This is because they ignore the correlations among process variables. The likelihood of encountering correlated variables has increased in recent years, because nowadays hundreds or thousands of variables are measured by process computers hooked up to a vast array of sensors and other analytical equipment. Moreover, final product quality data are today gathered at an ever increasing pace, though such variables are available on a less frequent basis than the process variables. All these variables should be analyzed together to extract the few underlying events which dominate the process. This is the point of departure for multivariate statistical process control, MSPC.

15.3.4 Transition from SPC to MSPC

The step from univariate to multivariate control charts is a very logical one, and, in fact, rather simple. Because the multivariate scores are mathematically orthogonal and optimal summaries of the measured process and quality variables, they are ideally suited for displaying in control charts. The scores are also less noisy than the original variables, since they are weighted averages. Thus, with multivariate control charts, all the process and quality variables are used together. The information residing in their correlation structure is extracted. This allows a more efficient tracking of the process over time for identifying anomalous process points, and for defining where the process is operating normally and when it is tending to depart from normal behavior [see e.g., Nijhuis, *et al.*, 1997; Teppola, *et al.*, 1998; Vanhatalo, 2009; Mirshel, *et al.*, 2013]. Multivariate control charts have been explained in detail in the literature [MacGregor and Kourti, 1995; MacGregor 1996; Wikström, *et al.*, 1998a; Gauchi and Chagnon, 2001], and will only briefly be reviewed here.

In the following sections, the aim is to outline some traditional control charts and their multivariate extensions. The most commonly used control charts are the Shewhart, CuSum and EWMA charts. For each chart type addressed, an explanation will be given together with illustrations drawn from the CUPRUM dataset.

15.4 The Shewhart chart

15.4.1 Construction in univariate application

The Shewhart chart represents the basic plotting architecture in SPC. A Shewhart chart is constructed simply by plotting the appropriate quantity versus time, together with indications of target and lower and upper control limits. Unless externally supplied, the target value is often set equal to the variable average. Similarly, if not otherwise stated, the lower and upper control limits are set as the average ± 3 standard deviations. With a Shewhart chart it is possible to detect large changes, like a shift in average or a general drift over time. A Shewhart chart is not good at finding small process shifts or drifts. Of course, large periodic changes may also be revealed by a Shewhart chart.

Figure 15.3 is of the Shewhart type and displays how the level of bismuth changes as a function of time. Clearly, in this variable the upper control limit is exceeded several times, which indicates unsatisfactory short-term process fluctuation. In this case, the numerical values cannot pass below the given lower control limit, because the amount of bismuth in copper cannot be negative. There is also a tendency for a weak drift towards higher numerical values (more impurity) towards the end of the sampling period.

15.4.2 Multivariate Shewhart charts

As mentioned, the PCA- or PLS scores, t, and Hotelling's T^2 based on such scores, can be plotted in a Shewhart chart. This is the key step towards conducting MSPC. Shewhart charts of scores will show how the process evolves over time in the respective PCA- or PLS model dimension. A target value can be included in the score control chart, together with warning and control (action) limits.

Figure 15.5 shows a Shewhart chart of the first score of the CUPRUM PCA model. We recall from Section 9.4 that this score accounts for 42% of the total variation in impurity pattern. Because the t_1 score is a weighted summary of the eight original variables – it can be thought of as an average trajectory of the eight individual quality variables – it is less prone to contain extreme numerical values than the individual variables. This is the reason why the process variation patterned by Figure 15.5 is milder in nature than the variation exhibited by the bismuth variable in Figure 15.3. In the first PCA model dimension, the process is stable and stays within control limits, except for a few extreme copper samples. We are not able to detect any strong time dependence. The same can be said for the second model dimension (t_2), but this plot is not shown for reasons of brevity.

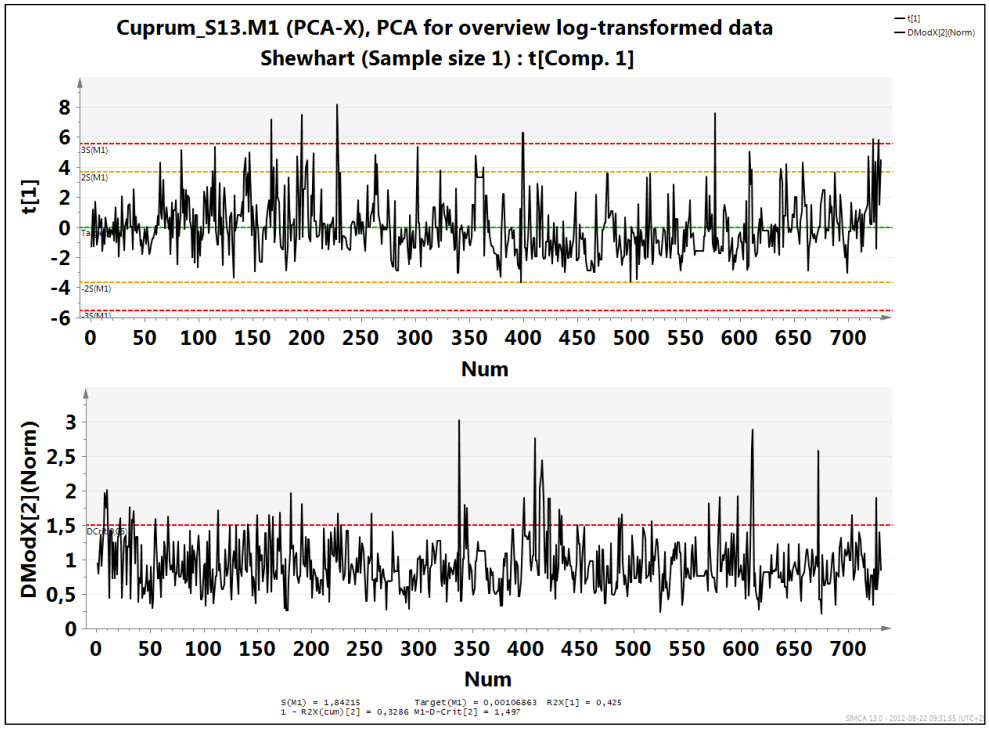


Figure 15.5: A Shewhart chart of the t_1 score of the PCA model for the CUPRUM dataset. The y-axis is the score value, and the x-axis represents sample number. The Shewhart plot of t_1 summarizes the systematic variation in the data. Tiled underneath is a plot of DModX, which enables an overview of the unsystematic process variation. Observe that the DModX control chart relates to the two-component PCA model.

Moreover, Figure 15.5 exemplifies how the unmodeled process variation (reflected by DModX) can be considered when a DModX-chart is tiled underneath the Shewhart chart. This plotting technique is very informative, and enables strong and moderate process outliers to be considered simultaneously. By inspecting these two control charts we get a good process status visualization of the copper manufacturing process. We can see if the copper quality tends to go out of control occasionally or if a more persistent trend is establishing itself. In the current application, no such conspicuous sequences of observations can be seen.

However, the Shewhart chart of t_1 does not convey the whole picture of how the copper process changes during one year of production. It is necessary to consider the other score, as well. Another, more appealing, alternative might be to construct a Shewhart chart of Hotelling's T^2. T^2 is a weighted summary of all scores, and gives an estimate of how far away an observation is from the center of a PCA- or PLS model hyperplane. It makes it possible to detect strong deviations in the systematic part of the data (strong outliers). Warning and action limits are here usually set at the 95% and 99% levels, respectively.

Figure 15.6 shows the T^2-statistic of the CUPRUM model plotted in a Shewhart chart. This plot is an overview of the process, considering all quality variables at the same time. We can see that the copper quality has gone substantially out of the nominal area in at least six (maybe even ten) instances.

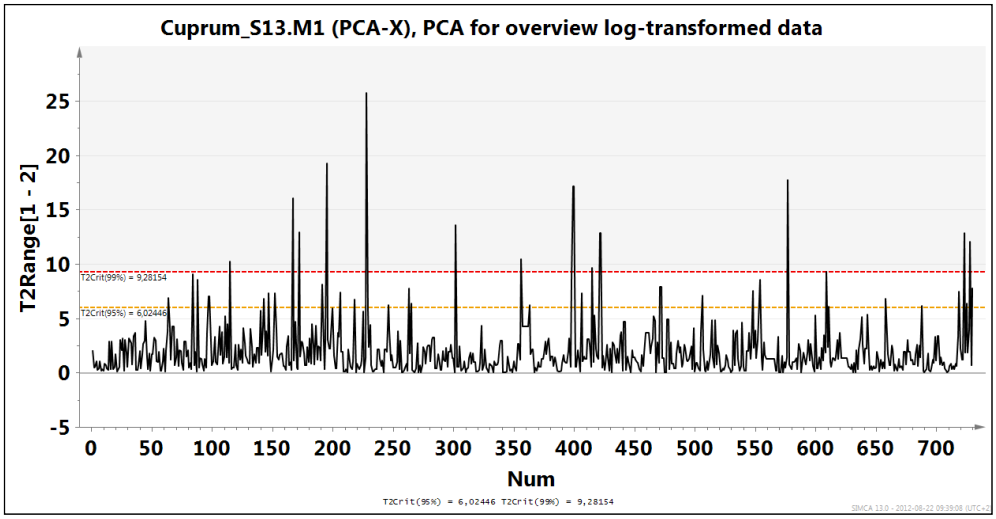


Figure 15.6: Shewhart chart with T^2 versus time order. Tolerance levels corresponding to 95% and 99% are given for T^2. An interpretation of this plot and the underlying model was given in Section 9.4.5.

15.4.3 Subgrouping

Subgrouping of data is often employed in conjunction with Shewhart charts. This technique is used when there is a rationale for doing so, e.g., subgrouping of observations over a day, a week, or a month. Subgrouping has the advantage of making the subgroup averages more nearly normally distributed, which, in turn, makes the control charts smoother and less susceptible to occasional extreme measurements.

By employing subgrouping, new types of control charts, here labeled as *Xbar-R* and *Xbar-s* become available. These new charts are actually a combination of two plots, one arising from the "X-bar"-part, the subgroup average, and another arising from either the "R"-part (subgroup *r*ange) or the "s"-part (subgroup *s*tandard deviation). A range chart is often set up when using small subgroups, and a spread (standard deviation) chart when utilizing a large subgroup size.

Figures 15.7 and 15.8 show control charts of score t_1, resulting from subgroup sizes 2 and 14, respectively. A subgroup of size 2 is justified since two copper samples were taken each day, and subgroup size 14 is linked to weekly averaging. In both plots, the upper X-bar (average) chart enables us to keep track of the variation in average from one day or week to another. The corresponding lower R(ange)- and s(pread; standard deviation)-charts show the variation within the subgroups.

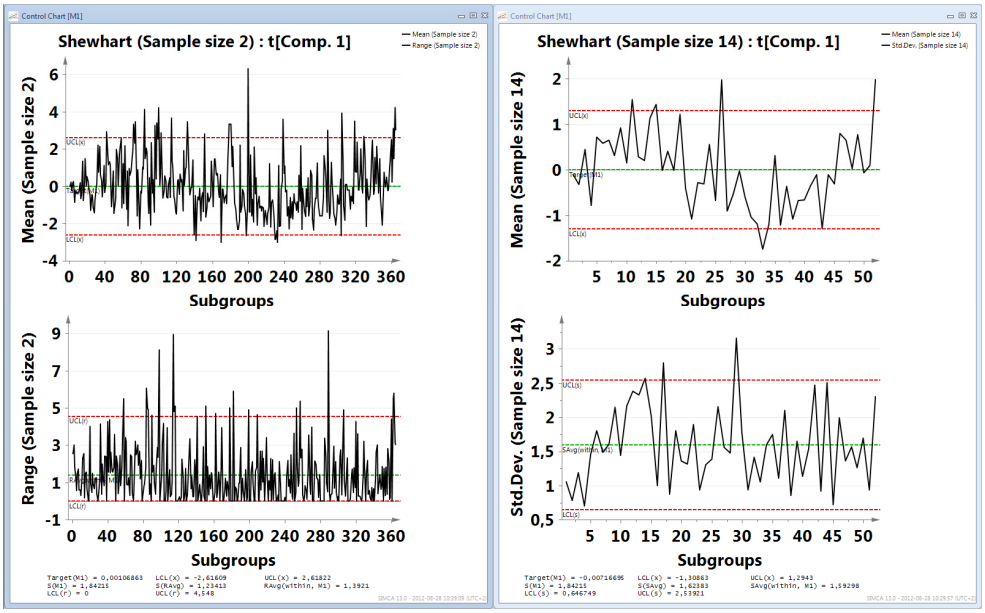


Figure 15.7: (left) Shewhart chart of t_1 of CUPRUM dataset using subgroup size 2. Lower part is a chart describing numerical range variation inside each subgroup. The horizontal direction corresponds to day number. Figure 15.8: (right) Shewhart chart of t_1 of CUPRUM dataset using subgroup size 14. Lower part is a chart reflecting variation in standard deviation in each subgroup. The horizontal direction corresponds to week number.

According to Figure 15.7, the most deviating (and highest) day average occurs at day 200, whereas the largest range variation is found during days 98, 114, and 289. From Figure 15.8 we can see that the variation across different weeks is much less pronounced than day-to-day variation. Only four weeks exhibit average values going outside the control limits, i.e., weeks 11, 26, 33, and 52. In the corresponding s-chart it is primarily the standard deviation of week 29 that is much larger than the SDs for the other weeks.

In summary, the few examples shown here demonstrate that by using rational subgrouping one can highlight information that is otherwise difficult to deduce from standard control charts. It is possible to maximize between subgroup variation and minimize within subgroup variation. Subgrouping is also applicable for other types of control charts, such as, CuSum and EWMA.

15.5 The CuSum chart

15.5.1 A cumulative sum of differences between target and measured values

In contrast to the Shewhart chart, the CuSum chart uses all of the observations since the last special event in order to uncover a new special event. A CuSum chart is useful for probing weaker process disturbances. The name of the CuSum chart derives from *cum*ulative *sum* of differences between the observed value and the target value. Its basic feature is that cumulative deviations from the target value, rather than individual numerical measurements, are charted over time.

15.5.2 Ways of implementing CuSum

There are different ways of constructing a CuSum chart. The classical approach is based on computing only one cumulative sum, i.e., the cumulative deviation (S_T) of all observations from target (Figure 15.9). In many cases, the CuSum can wander remarkably far in a control chart. Therefore, in SPC applications, the CuSum chart is often accompanied by a V-mask, which will aid the detection of unwanted process behavior (Figure 15.10). We will not go into details regarding the use of the V-mask; details can be found in the literature [Massart, *et al.*, 1998; Rius, *et al.*, 1998]. The V-mask is positioned at a fixed distance, d, from the last

plotted point. This lead distance, d, depends in part on the scaling of the chart axes. When the CuSum line cuts one of the action limits of the mask, this constitutes a signal that the process needs attention.

$$S_T = \sum_{t=1}^{T} (y_t - \text{target})$$

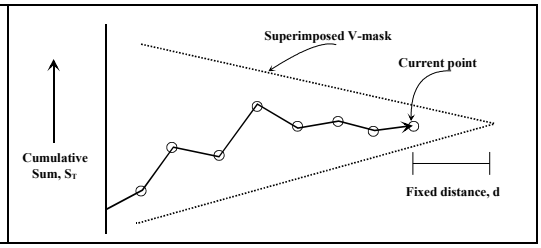


Figure 15.9: (left) Computation of the cumulative sum of differences between observed values and the process target. Figure 15.10: (right) Illustration of principle of V-mask. When the CuSum line intersects with one of the "legs" of the V-mask, action is needed to reset the process.

A more common way of presenting a CuSum consists of subgrouping the data and calculating more than one cumulative sum of deviations from target. One cumulative sum adds the differences on the high side (values/scores larger than the target value), and the other cumulative sum adds the differences on the low side (values/scores smaller than the target value). In the plotting, three points are plotted for each subgroup, one point is the deviation from the target, the second represents the high side sum, and the third the low side sum. This is a way of increasing the diagnostic power of the CuSum approach. If, for instance, the series has been showing positive deviations for a while, the positive sum has been increasing and the negative sum decreasing or stopping at zero.

15.5.3 Multivariate CuSum charts

The CuSum approach which involves the computation of more than one sum is the one implemented in SIMCA. With this technique, univariate CuSum charts can be obtained by plotting individual variables, and multivariate charts by displaying model scores. Figure 15.11 gives a CuSum chart of CUPRUM score t_1 using subgroup size 2. The information in this graph is somewhat different from Figures 15.6 – 15.8. The CuSum chart indicates the existence of a long series of consecutive negative deviations from the process average, commencing at day 210 and ending at day 320. This information was not easily retrievable from the previous Shewhart charts.

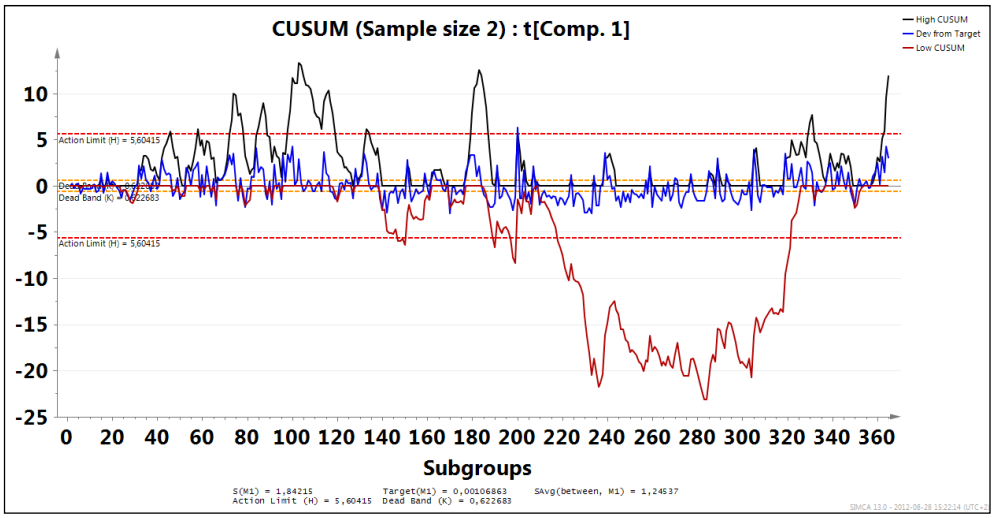


Figure 15.11. Multivariate CuSum control chart for t_1, and using subgroup size = 2. This subgrouping means that the x-axis corresponds to 365 days. The plot shows two cumulative sums, high and low (see text), and the deviation from the target. In the plot, the lines "K" represent the area where the process is operating on target, whereas the lines "H" designate action limits.

15.6 The EWMA chart

15.6.1 An approach towards exploring process memory

The last type of control chart that we will describe is the exponentially weighted moving average (EWMA) control chart [Hunter, 1986]. Usually, the EWMA approach is applied to model a dynamic process with memory and drift. The basic idea with EWMA is to model a variable as a weighted moving average, with recent observations weighted more heavily than older observations. Exponentially decreasing weights

$$v_i = \lambda(1 - \lambda)^{t - i} \qquad \text{(eqn. 15.1)}$$

are used for the i^{th} observation preceding the current one ($i = t$). A low value of λ (close to 0) corresponds to a long memory of the process and a high value of λ (close to 1) to a short memory (Figure 15.12). The value of λ may be estimated from the data as the one minimizing the predictive error sum of squares, or defined by the data analyst. If user specified, λ is typically set to 0.2 ± 0.1.

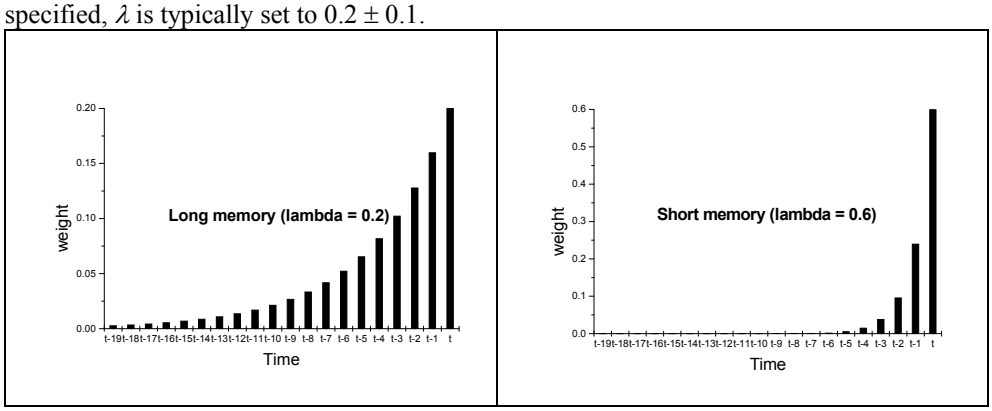


Figure 15.12: Schematic illustration of how the EWMA memory changes with changing values of λ (lambda). Long memory is shown in the left-hand part, and short memory in the right-hand part of the picture.

15.6.2 Ways of implementing EWMA

The EWMA approach can be implemented in two versions, predictive and as a first-order filter. Predictive EWMA is given by

$$\hat{x}_{t+1} = \hat{x}_t + \lambda(x_t - \hat{x}_t) \qquad \text{(eqn. 15.2)}$$

and the first-order filter EWMA by

$$\hat{x}_t = (1 - \lambda)\hat{x}_{t-1} + \lambda x_t \qquad \text{(eqn. 15.3).}$$

Here a "hat" value corresponds to a predicted value. In addition, one often uses subgrouping of the data to stabilize the modeling. When the subgroup size is 1, the process standard deviation is estimated from the data, converted to upper and lower control limits, and displayed alongside the EWMA estimates. As mentioned before, the subgroup size should reflect a natural sub-division of the observations. For instance with the copper data, two samples were taken each day and the logical subgroup size is consequently 2. With subgroup size > 1 the process standard deviation is estimated as the mean square successive difference, i.e., the between subgroup variation.

15.6.3 Multivariate EWMA charts

Multivariate EWMA charts are obtained by plotting PCA and PLS model scores. Their univariate counterparts can be created by plotting individual variables, but no such plots are given here.

In the copper electrolysis process each cathode was left in the electrolyte for 10 days and each anode for 20 days. This means that there may exist a memory pattern in the data. This can be checked with the EWMA approach. Figure 15.13 displays the multivariate EWMA chart for t_1. Note the use of the subgroup size 2, because two copper samples were analyzed

every day. The memory parameter $\lambda = 0.09$ was estimated from the data as the value minimizing the predictive error sum of squares of one-ahead forecasts.

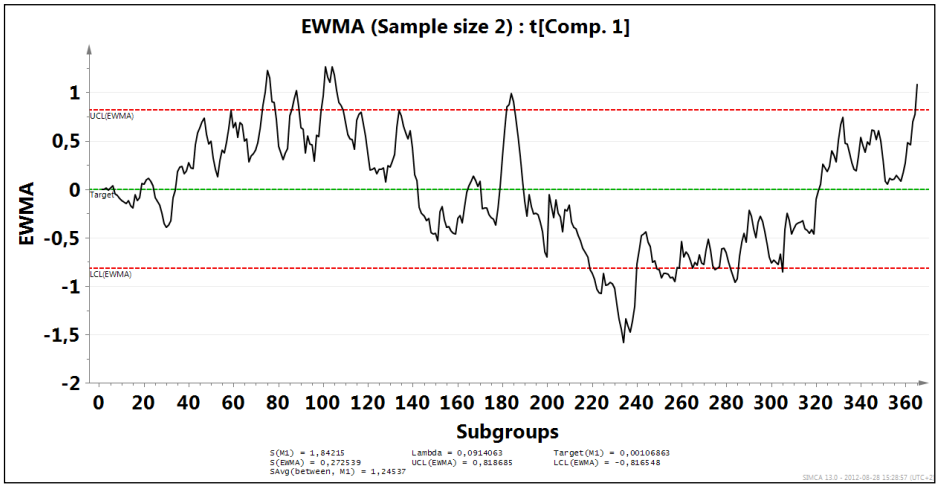


Figure 15.13: Multivariate EWMA control chart for t_1, and using subgroup size = 2. Lambda = 0.09 was estimated from the data. The subgrouping used means that the x-axis corresponds to 365 days. Upper and lower control limits (UCL, LCL) and the target (estimated as the mean process observation) are given.

As shown by Figure 15.13, an EWMA chart is smoother than the corresponding Shewhart chart (Figure 15.7) and therefore capable of detecting small process shifts and time dependencies. Evidently, there are characteristic time trends in the CUPRUM data, which were not so apparent in the Shewhart control charts. To better understand these time series properties one must (i) identify the appropriate family of time series models, and (ii) estimate the model parameters. Such multivariate time-series analysis, coupled to lagging of latent variables, is discussed in Chapter 17.

Moreover, when comparing the EWMA (Figure 15.13) and CuSum (Figure 15.11) approaches to monitoring the CUPRUM process, we realize that similar features are discernible in both control charts. However, the CuSum chart shows more clearly the long continuous period of negative deviations from the process average, commencing at day 210 and ending at day 320. Interestingly, the lower control limit in Figure 15.13 and the action limit in Figure 15.11 would both issue an alarm around day 219.

15.6.4 EWMA and its relation to Shewhart and CuSum charts

Above, we explored the extent of process memory in the copper quality data, by using the exponentially weighted moving average approach. EWMA contains an adjustable parameter, λ. It may be estimated from the data (as the value resulting in the best one-ahead predictions) or set by the user. Lambda = 1 implies no memory (Shewhart chart) and lambda = 0 corresponds to infinite memory (CuSum chart). Hence, the EWMA control chart can be seen as located between the Shewhart and CuSum charts. The estimated value of $\lambda = 0.09$ indicates a comparatively long memory.

Furthermore, this small λ-value of EWMA means that multivariate CuSum control charts may also be appropriate (see Figure 15.11). In the CUPRUM dataset, a relatively long period of systematically negative deviations from the process mean was easier to detect with CuSum than with EWMA. Such a deviation from target is vital information for a process operator or a quality engineer. According to our experience, the use of a multivariate CuSum chart is recommended when EWMA has a low value of λ, say, below 0.15.

15.7 Example: PROC1A

15.7.1 A score scatter plot shows direction of change

In Sections 15.4 – 15.6 the basic approaches to control charting in process analysis were introduced and exemplified using the CUPRUM dataset. In the current section, we will put into practice what we have learned about MSPC, with reference to the PROC1A dataset. The PROC1A dataset was analyzed in Chapter 9, and described in the introduction to the current chapter.

It has been established that the process began to behave strangely after time point 70 and went out of control after time point 80. Figure 15.14 shows a t_1/t_2 score plot of the three-component PCA model (cf. Section 9.5) representing good past process behavior. When applying the obtained model to the complete dataset (92 observations) the picture seen in Figure 15.15 was acquired. An instantaneous warning that the process tends to depart from normal behavior after time point 80 is obtained.

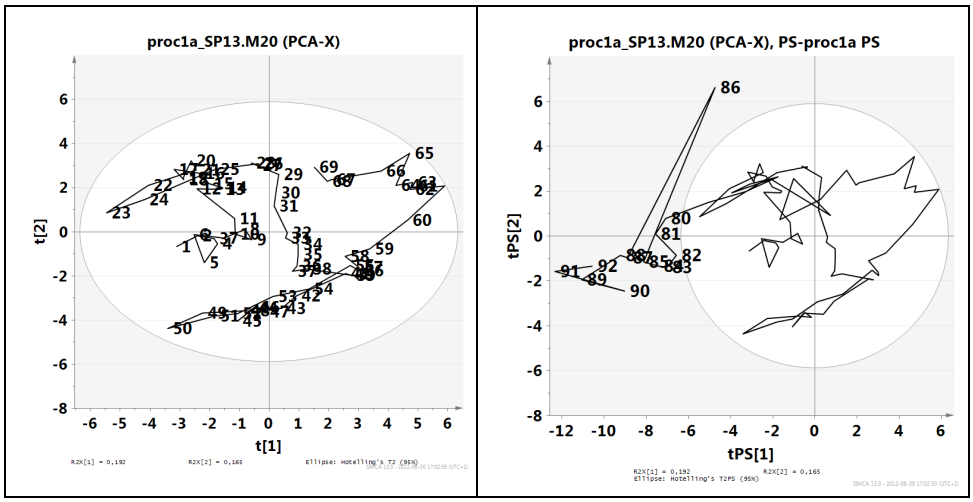


Figure 15.14: (left) PCA t_1/t_2 score plot of the PROC1A dataset. The model was founded on the first 69 observations, representing good past process operating conditions. Figure 15.15: (right) When applying the obtained PCA model to the entire set of 92 observations, the shift in process behavior after time point 80 is obvious.

The nice feature of the scatter plot representation is that one can discern a *direction of process change* in relation to the normal operating area. The PROC1A process drifts away in the "southwest" direction in the t_1/t_2 score space. By coupling this information with loading and contribution plots, it is possible to interpret which variables participate in expressing the observed process drift. This was discussed in Section 9.5.

15.7.2 A multivariate T^2 control chart shows magnitude of deviation (but not direction)

However tractable the scatter plot representation of scores may seem, it does not convey the whole story of what is going on in the PROC1A dataset. Because there are three principal components in the model, we must interpret all of them to get a detailed understanding of how this process evolves over time. A convenient way of overviewing the three PCs is offered by Hotelling's T^2, which is a measure of how far away from the center of the projection an observation is situated. The problem with this alternative representation is that it is a squared function of the PCA scores and only gives the absolute magnitude of a process deviation. Hence, the sense of direction in which the process is moving is lost.

Figure 15.16 shows a multivariate Shewhart chart of Hotelling's T^2 after three PCs. In this graph, the warning limit (T^2_{crit} 95%) is represented by the yellow line and the action limit (T^2_{crit} 99%) by the red line. The warning limit has an interpretation similar to the ellipse in

the score plot of Figure 15.15, although these two limits do not relate to the same model complexity. Apparently, the warning limit in Figure 15.16 and the tolerance ellipse in Figure 15.15 would both issue a warning that something strange is going on in the process at the same time point (80). The action limit in the T^2-chart would alert the operator after time point 85. However, because of its absolute nature, the T^2-chart does not state *how* the problematic process period relates to the good past behavior.

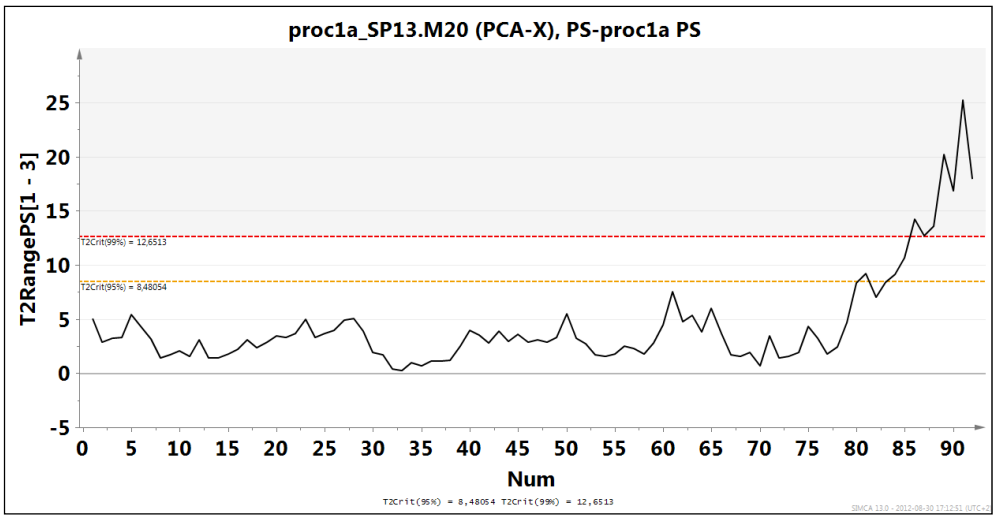


Figure 15.16: Multivariate Shewhart chart of Hotelling's T^2 of the PROC1A dataset after three components.

15.7.3 Important scores can be charted alone

Sometimes it is of interest to chart PCA scores independently of one another. This is of particular relevance when a certain score has been interpreted as reflecting a fundamental process property, which is of great interest as far as process regulation is concerned. For instance, in the PROC1A model, the first score is a good summary of the variables x1in, x3in, xemd, xgnx, xoen, and xpen (no loading plot given here). Assuming this group of variables to be heavily inter-related and jointly reflecting a fundamental process property, say, internal reactor environment, it might be warranted to follow the t_1 trajectory over time.

Figure 15.17 shows a combined Shewhart and EWMA control chart of t_1. According to this plot, only mild process changes occur in the first model dimension. However, it is evident that the process drifts away towards the end of the investigation period. Interestingly, the Shewhart and EWMA charts are extremely similar. The one-ahead forward shift of the EWMA curve arises because EWMA is here used in predictive mode. Had we chosen to use EWMA as a first order filter instead, the EWMA line would have been placed on top of the Shewhart line. Here, λ has been estimated at one (1). This indicates that there is no process memory. Hence, the EWMA and Shewhart charts should look similar when co-charted in the same graph.

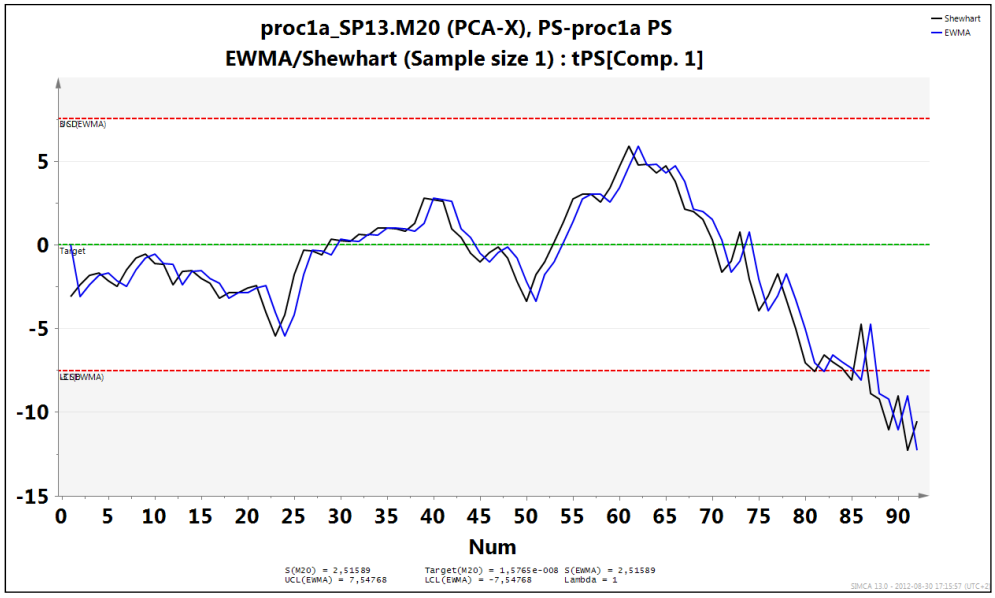


Figure 15.17: Multivariate Shewhart and EWMA charts of t_1 score of the PROC1A dataset. The lambda parameter has been estimated from the data and found to be unity. This suggests that no memory exists in the process, and therefore the Shewhart chart is applicable. This also means that the Shewhart and EWMA charts should be very similar.

15.7.4 Real-time process monitoring using SIMCA-online

A driving force behind creating a multivariate model of a process is to deploy it to track the performance of the corresponding process in real-time. As we shall see in this section, this is possible through the use of the SIMCA-online software. SIMCA-online executes one or more models -- developed in SIMCA offline -- in a real-time environment and thus enables referencing current and future process performance against past good behavior. Early fault detection, root-cause diagnosis and on-line quality control are the benefits, items which also are the keys to an efficient process.

SIMCA-online offers the same plethora of control charts and diagnostics as SIMCA offline. Below, we illustrate a few options using the PROC1A dataset. The process model in use is the three-component PCA model discussed in Sections 9.5 and 15.7.1 defining good operating conditions, which is executed on the additional data not used for model building. Since the model contains three principal components it is reasonable to use the T2Range Shewhart chart shown in Figure 15.18. We recall that the basics of Hotelling´s T2 was discussed in Sections 3.5 and 9.4 and the basics of the Shewhart chart in Section 15.4.2.

As shown by Figure 15.18, SIMCA-online issues warnings and alarms, depending on the warning and alarm rules that have been configured for this particular multivariate model. The black arrow inserted into Figure 15.18 highlights a time-point at which SIMCA-online has issued an alarm to alert the process engineer that something is not right with the process. Using drill-down technology (see example below), the process engineer can quickly interpret the process deviation and decide whether the alarm should be acted upon, or not.

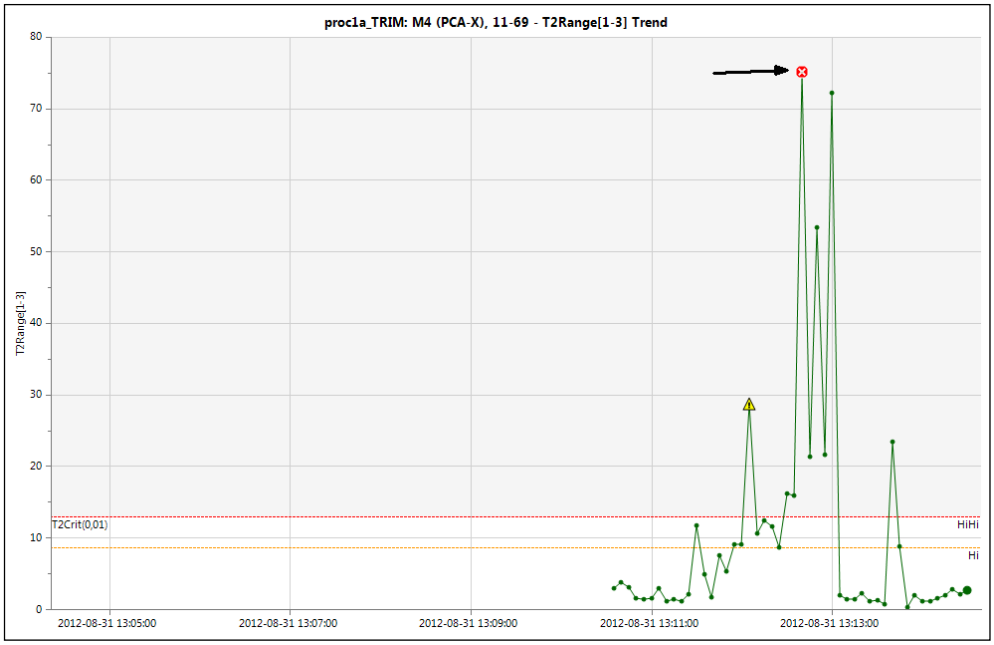


Figure 15.18: Multivariate Shewhart chart of T2Range of the model for the PROC1A dataset. The ability of SIMCA-online to issue warning and alarm messages is illustrated. The warning is indicated by the triangle and the alarm by the combined cross-dot symbol.

Another powerful display is the DModX control chart, which covers the orthogonal complement to Hotelling´s T2 (see Figure 15.19). The arrow in Figure 15.19 points to the same time point as was previously detected in Figure 15.18. It should be noted that an observation (time point) can be off in either T2 or DModX, or in both parameters at the same time. We recall from the discussion in Chapter 3 that a deviator in score space or Hotelling´s T2 is usually a strong outlier, whereas a deviator in DModX has lower urgency and is usually a weaker outlier.

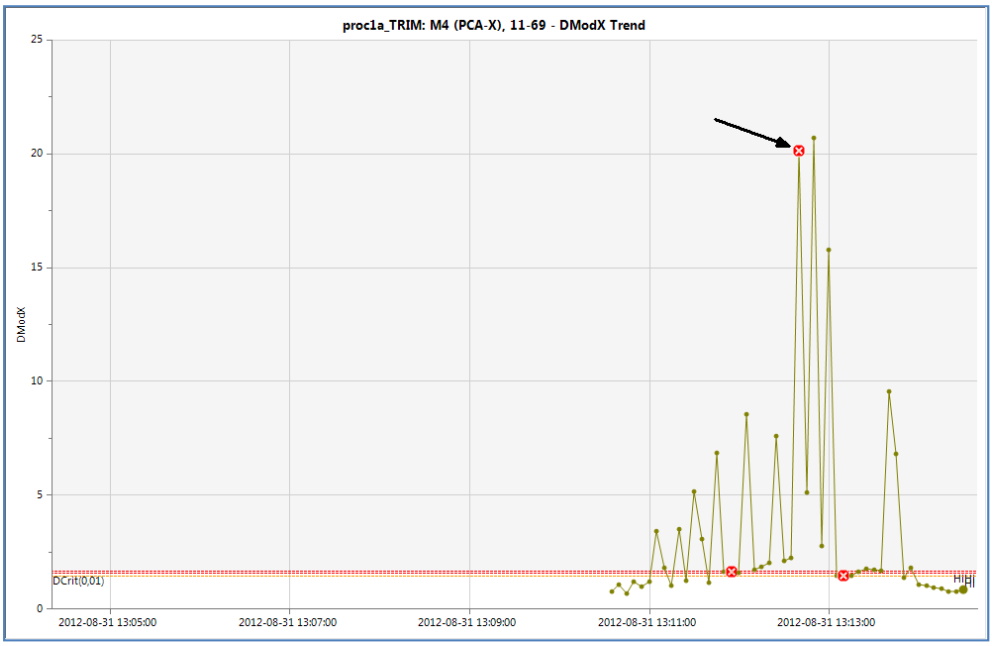


Figure 15.19: Multivariate Shewhart chart of DModX of the model for PROC1A. The arrow points to a time point that was previously identified in Figure 15.18.

Regardless of deviator detection principle, interpretation of the root cause of a process upset is possible by using contribution plots. We recall that the basic principles of contribution plots were introduced in Section 9.4.4. Contribution plots are highly informative in process modeling and monitoring [Westerhuis, *et al.*, 2000; Undey, *et al.*, 2003; Vanhatalo, 2009]. By double clicking on the point close to the arrow in Figure 15.19, the contribution plot displayed in Figure 15.20 is automatically launched by SIMCA-online. It does not take long for the analyst to realize that the process variable x6in is much lower than anticipated from the correlation structure picked up by the PCA model. A second double click on the highlighted variable will then open the corresponding trend plot (Figure 15.21), which corroborates that the numerical value of x6in is way too low.

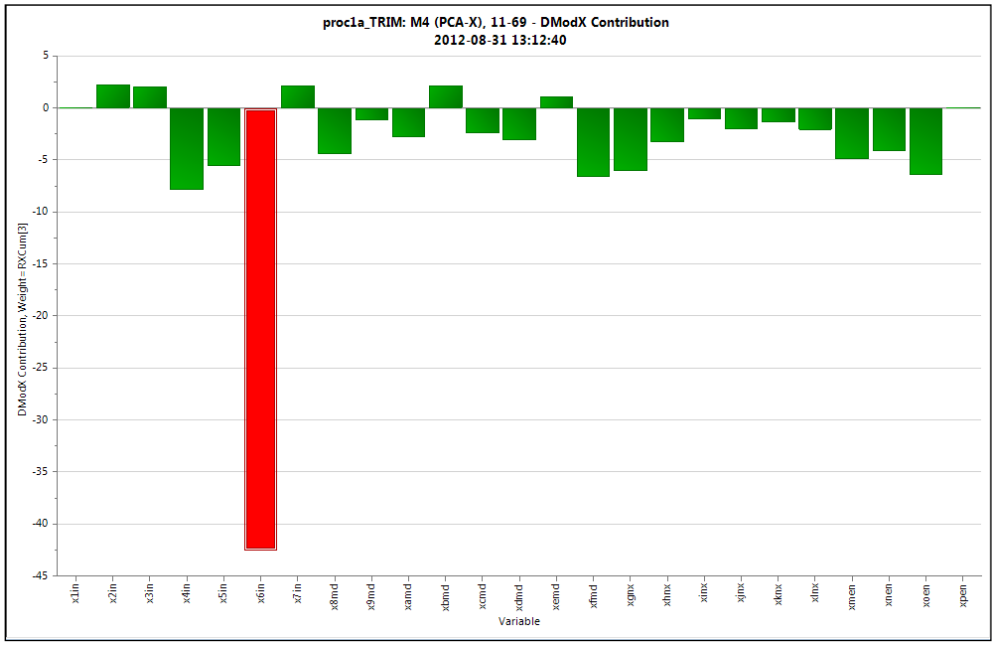


Figure 15.20: Contribution plot in DModX mode, corresponding to the time point highlighted in Figure 15.19. The contribution plot reveals that one of the process variables (x6in) is much lower than it should be. By double clicking on the highlighted variable its trend plot is instantaneously displayed, see Figure 15.21.

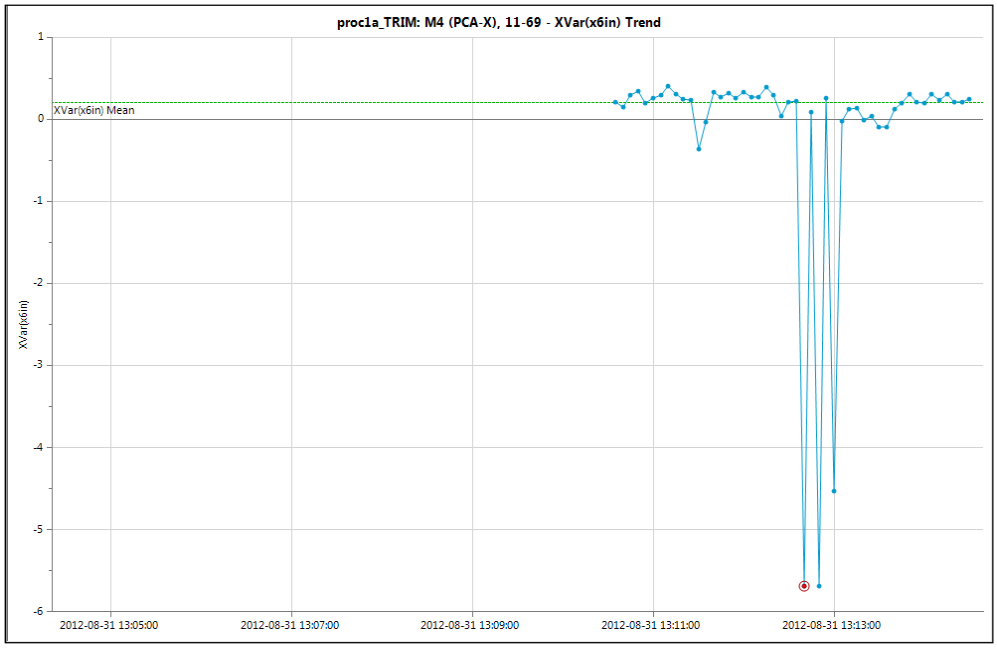


Figure 15.21: Trend plot of the process variable x6in. The trend plot shows that the reason for the process alarm in the DModX control chart (depicted by the arrow in Figure 15.19) is the x6in signal being extremely low.

15.7.5 Summary of MSPC of PROC1A dataset

With PCA, the process upset after time point 80 was readily identified. By the addition of MSPC a more formal assessment of the encountered deviation was possible. In general, in this kind of application it is possible to chart both scores and a summary of scores (T^2) over time, and use warning and action limits for decision-making. Similar possibilities hold true for DModX or any other multivariate parameter.

This example also shows how real-time process monitoring and early fault detection can be carried out using the SIMCA-online software. Generally, three implementations of a real-time process model can be discerned.

- Monitoring: The basic level is to use SIMCA-online for monitoring and early fault detection, which provides better use of equipment, time and starting materials. This promotes process understanding.
- Prediction: On this level, the additional goal is to predict the quality of intermediates and the final product, before the production is finished. The benefits are obvious both in terms of improved product quality and increased process knowledge.
- Control: On this level, the additional objective is to be able to accomplish automatic adjustments to the process, as it evolves, based on the multivariate signals output from a model. This could be said to be the ultimate goal, i.e., that during run-time you turn process upsets into controlled process efficiency.

15.8 Questions for Chapter 15

1. What are the intrinsic problems of process data, which can complicate the data analysis?
2. What is SPC?
3. What is the basic assumption underlying SPC?
4. What is a special event?
5. What is MSPC?
6. Which multivariate parameters are normally charted in MSPC? What do they mean?
7. When is the Shewhart chart useful?
8. What is subgrouping, and when is this technique applicable?
9. When is the CuSum chart useful? The EWMA chart?
10. How is EWMA "positioned" in relation to the Shewhart and CuSum plotting approaches?

15.9 Summary and discussion

15.9.1 The basic control charts

Statistical process control (SPC) is the classical approach for investigating stability and reliability of industrial processes over time. The Shewhart, CuSum and EWMA control charts are the basic tools used for process monitoring. The Shewhart chart is the simplest and most straightforward. This chart allows the mean level and the variability of the process to be charted over time. It can be used for individual measurements and also for data arising from subgrouping. Subgrouping is used to make the charted data approach normality, which has the advantage that the resulting control charts often become smoother.

The Shewhart chart is useful for detecting large and sudden shifts in a process. However, the main drawback of the Shewhart chart in relation to other control charts is its poor ability

to detect subtle process fluctuations. The CuSum and EWMA charting approaches are more efficient for tracking small process changes. This feature arises because in a Shewhart chart, the last plotted point depends only on the last measurement, while in a CuSum chart it depends on all previous observations. In an EWMA chart, the last plotted point depends on the value of the adjustable memory parameter λ.

The EWMA statistic can be used in two ways, as a first-order filter and in predictive mode for predictions of one observation ahead (at time $t + 1$). Of course, predictions can be made beyond $t + 1$, at time $t + 2$, $t + 3$, etc., but such estimates are less reliable. When applying EWMA, the memory parameter λ must be specified. In SIMCA it is possible to estimate λ from the data, but it may also be specified by the user. Usually, we find that λ is set at 0.2 ± 0.1. Note that with a low value of λ, say below 0.15, the CuSum charting approach becomes applicable.

15.9.2 Control chart limits

The classical control charts, univariate as well as multivariate, contain limits that assist the process engineer in taking adequate decisions. Usually, such control chart limits are asymptotic in nature and therefore require statistical theories for their estimation [Babamoradi, et al., 2013]. Asymptotic control limits work well for process datasets that are homogeneous, representative and approximately normally distributed. In this context, homogeneity means fairly uniformly distributed process data points in e.g. a score scatter plot. A score plot with strong clustering, e.g., with a few "islands" of densely occurring process points and much "water" surrounding them, would not correspond to a homogeneous process dataset.

Realizing the need for alternative and more flexible limits for control charting purposes, Babamoradi and coworkers carried out an elegant and systematic exploration of control limits applicable in MSPC where data are both grouped and non-uniformly distributed [Babamoradi, et al., 2013]. Their approach is based on bootstrapping with replacement and several interesting ideas are put forward, including a novel method for computing confidence limits for contribution plots.

15.9.3 Approaches to MSPC

In this chapter we have emphasized the natural transition from SPC to MSPC. Simply by selecting to chart a multivariate parameter resulting from a projection method (PCA, PLS, ..), we accomplish MSPC. Usually, this leads to a better understanding and overview of process behavior, than by tracking individual variables in univariate SPC charts. The benefits of the multivariate approach to SPC have been discussed at length in several papers by the process analysis research group at McMaster University, Hamilton, Ontario, Canada, a research group centered around professor John MacGregor [see for example MacGregor and Nomikos, 1992; MacGregor and Kourti, 1995; MacGregor, 1996].

Of course, there exist other approaches to MSPC. Often process data are not two-way, but multi-way or hierarchical in nature. Multi-way data occur for instance in batch modeling, in process applications where hyphenated analysis techniques (GC-MS, LC-MS, ..) are employed, and in image analysis and remote sensing. Hierarchical data occur when processes are divided into conceptually meaningful process blocks, which may be modeled separately or in conjunction. A description of multi-way PCA and PLS models can be found in [Wold, *et al.*, 1987], and a description of the application of hierarchical modeling to process data in [Wold, *et al.*, 1996]. An in-depth discussion regarding hierarchical modeling is given in Chapter 19.

In order to address multi-way data one may either unfold the matrix structure such that it becomes a two-way table, or apply other data analytical models which assume multi-way or hierarchical data. As we shall see in the next chapter (Chapter 16), unfolding works very well when conducting MSPC on batch data. We call this approach batch statistical process control, BSPC. However, there are also other multi-way techniques, such as the PARAFAC model, which may be applied to multi-way process data. Bro has published a good account

of the three-way PARAFAC approach, and used process fluorescence measurements from the white sugar manufacturing industry to illustrate the concepts and benefits of PARAFAC [Bro, 1999]. Wise and co-authors applied an extension called PARAFAC2 to batch data [Wise, *et al.*, 2001].

Furthermore, Rius and coworkers have published a good tutorial article describing how artificial neural networks and recurrent neural networks may be used in MSPC [Rius, *et al.*, 1998]. Their discussion is focused on how to apply SPC and MSPC in analytical process applications. As pointed out by Rius and coworkers, the limitations of the neural network approach are that long training times are sometimes needed, and that such models are generally difficult to interpret.

15.9.4 SPC/MSPC and relation to feedback process control

The basic assumption of SPC is that a process is in a state of statistical control unless a special event occurs. In principle, this means that it is assumed that the process mean and standard deviation are almost time invariant, until a departure from normal operating conditions takes place. Usually, process operators try to maintain normal process behavior to the maximum possible extent, by adjusting key regulatory process variables. As a consequence, a set of process data representing normal behavior often display a low degree of variation. When variation in such data are subjected to PCA or PLS, there is little information to found a model on. This often results in models with comparatively low R^2- and Q^2 values.

Furthermore, we have to consider the impact of automatic feedback process controllers, not just process operator actions. Automatic feedback control is applied to reduce variability in process and product variables. A feedback control system compensates for known or expected disturbances in important process and product variables by adjusting other variables. In this way, variability in important variables is often transferred to the less important manipulated variables.

According to MacGregor and coworkers [MacGregor, *et al.*, 1995], "SPC methods should be applied on top of the process and its automatic control system in order to detect process behavior that indicates the occurrence of a special event. By diagnosing causes for the event and removing them (rather than simply continuing to compensate for them), the process is improved."

15.9.5 Alternative approaches to process monitoring and control

We will now address some additional process monitoring techniques other than SPC/MSPC. One model-based approach to process monitoring is called *state estimation*. State estimation techniques use theoretical or mechanistic models for process monitoring, which usually take the form of a set of differential equations [MacGregor and Nomikos, 1992]. The objective of state estimation is to predict the internal state from a small set of sampled and noise-corrupted measurements. The estimation of process state is accomplished by observers or Kalman filters, which reconstruct the various states. Parallel filters are usually used for different fault states.

There are several requirements that need to be fulfilled before applying state estimation. One must have access to a good mechanistic model that reflects the main physical and chemical phenomena occurring in a process. One must also have access to good sensors that provide relevant and precise information about the state of the process. And, of course, it is crucial that the process analyst has good insight into what is being modeled, e.g., known disturbances, faults that may occur, and other varying conditions.

In a work by Hartnett and co-authors it was shown how PCA could be incorporated into a state estimation framework [Hartnett, *et al.*, 1999]. They demonstrated that PCA could be used to build a predictive model according to a state estimation format, and that PCA would facilitate some of the initial modeling requirements.

Another approach to process monitoring is based on the establishment of expert systems. An expert system often includes a set of qualitative and quantitative governing descriptions

arranged in a hierarchical structure. Such a governing description can be a qualitative model based on good process knowledge, detailed information and causal description of the process, and lots of representative process data on occurring process faults. In general, an expert system is composed of a set of logical production rules (e.g., IF *symptom* AND *symptom* THEN *conclusion/action*), where each conclusion may serve as a symptom in the next governing rule until the final diagnosis or decision is reached.

Finally, an approach of Stork and Kowalski will be highlighted. This approach is called the *redundant sensor voting system* (RSVS), and is intended to distinguish between process upsets and sensor malfunctions [Stork and Kowalski, 1999]. It is important to differentiate between these two types of process "events" in the monitoring of a process. Process upsets threaten quality of product and may require immediate intervention by the process operator, while sensor malfunction can be compensated for mathematically.

The RSVS approach is based on a set of probabilistic and empirical rules for distinguishing between sensor failures and process upsets. The basic working principle is as follows: In the presence of sensor redundancy, process disturbances will be registered by a multitude of sensors. Through the set of probabilistic and empirical rules, a threshold value is defined regarding how many sensors should be registering a certain type of process change. Then, the diagnosis is based on the fact that process upsets are typically registered by a band of correlated sensors, whereas a sensor failure is often localized to one or a few sensors. According to the RSVS procedure, if the number of identified sensors exceeds a certain threshold value, a process upset is inferred. On the other hand, if the number of identified sensors falls short of the reference value, this indicates sensor malfunction.

16 BSPC

16.1 Objective

The objective of Chapter 16 is to provide an introduction to the analysis of batch process data. In so doing, we shall familiarize ourselves with an approach called *batch statistical process control*, or BSPC for short. The chapter starts with an introductory discussion of certain hallmarks of batch data, and then describes the essential features of BSPC. The multivariate approach to BSPC is exemplified by two examples, one drawn from the production of yeast (BAKER'S YEAST), and the other from the semiconductor industry (WAFERS).

16.2 Introduction

16.2.1 A finite duration process

In the foregoing chapter (Chapter 15) methods for multivariate statistical process control (MSPC) of continuous processes were outlined. In this chapter, the aim is to describe analogous techniques which are tailored for batch statistical process control (BSPC). The development and application of methods for BSPC is of great practical significance since an abundance of industrial processes are of the batch type. Batch-wise manufacturing processes are found in the chemical, pharmaceutical, bio-technical and semi-conductor industrial sectors. Typical examples include baker's yeast production, beer brewing, car painting, PVC polymerization, wafer etching, and fermentations in pharma and biopharma.

In contrast to a continuous process, a batch process has a *finite duration*, from initialization to completion. Usually, trajectories of batch process variables describe dynamic time dependencies, that is, the numerical values of batch variables change as a result of batch progress (Figure 16.1). Figure 16.1 displays trajectories of seven monitored variables in the manufacturing of baker's yeast (further details are given in Section 16.3.1). Clearly, these change differently during batch evolution. The question is: which of these variables correlate with final product quality?

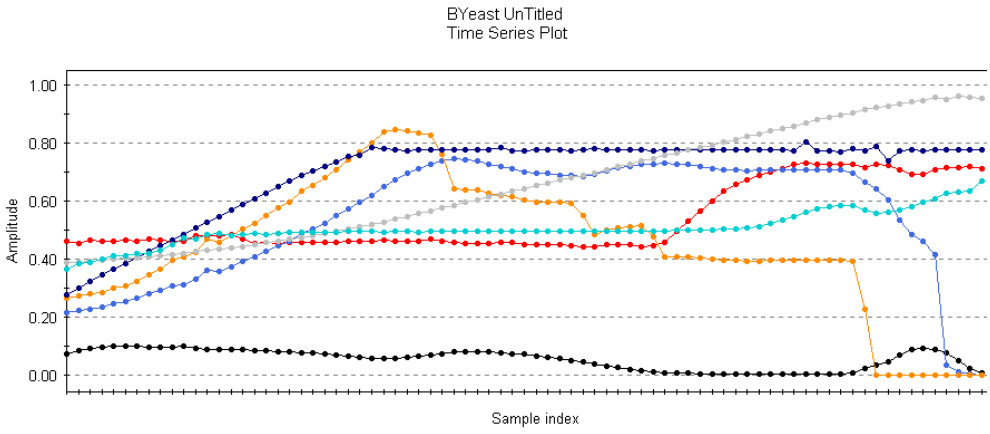


Figure 16.1: A batch process has a finite duration. The time series plot shows how some critical batch process variables change during the last phase of production of baker's yeast. Note that data have been scaled to fit into the same plot.

In addition to varying operating conditions within each batch, batch processes generally exhibit moderate to large batch-to-batch variation. Such between batch variation may arise because of deviations of the process variables from their target trajectories, errors in the charging and initialization of batches, or fluctuations linked to variation in starting materials and their impurity profiles. Unless such unwanted or atypical conditions are detected and corrected, low-quality product produced by one batch, or a whole sequence of batches, might have to be discarded.

16.2.2 Batch processes form three-way data tables

Batch processes give rise to data tables that are different from the two-way data structures which we studied in the earlier chapters. Measured data from batches are handled in three-way matrices (Figure 16.2). A batch starts by charging the reactor with starting material. After initialization, the observation of the first process point is made. This gives rise to a vector x_{i1} with K elements, x_{i1k}. Batch evolution of batch i is then registered by the measurement of the same set of K variables at time 2, 3, …, until batch completion time J. Hence, in a batch process where K variables are registered at frequent time intervals ($j = 1,2,...,J$), each batch results in a two-way data matrix of dimension $J \times K$. Consequently, a set of N normal batches then corresponds to a three-way data structure of dimension $N \times J \times K$ (Figure 16.2).

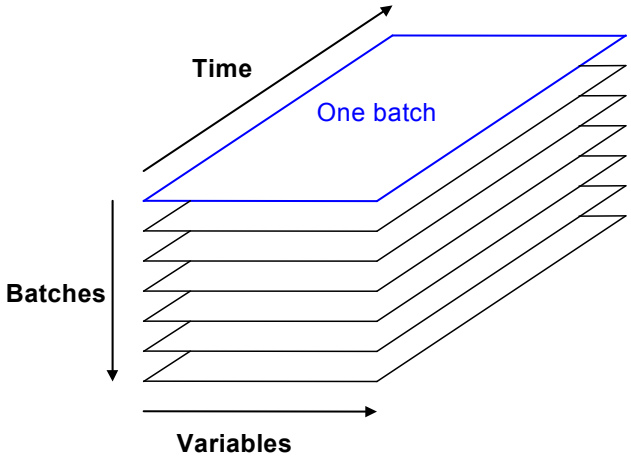


Figure 16.2: A three-way table of historical batch process data. The data table comprises N batches, J time points, and K variables.

We emphasize that the situation sketched in Figure 16.2 is an idealized picture of batch process data. Often there are missing data, giving rise to holes in the three-way matrix. Furthermore, in many cases batches have different lengths, i.e., the batches may extend differently in the time direction. Special alignments involving stretching and shrinking procedures are then needed. This is discussed in Section 16.8.5.

Furthermore, there might be more than one block of batch process data (Figure 16.3). Initial conditions data are given by one data table (often called the Z-matrix). Information of relevance here would be characteristics of raw material, environmental conditions like temperature, pressure, and humidity, and perhaps reactor-specific features, if many reactors are run in parallel for the same purpose. In the second data matrix, batch evolution data are gathered. This matrix is often called the X-matrix. Naturally, the last block of data, often called the Y-matrix, is composed of results and product quality data for each batch.

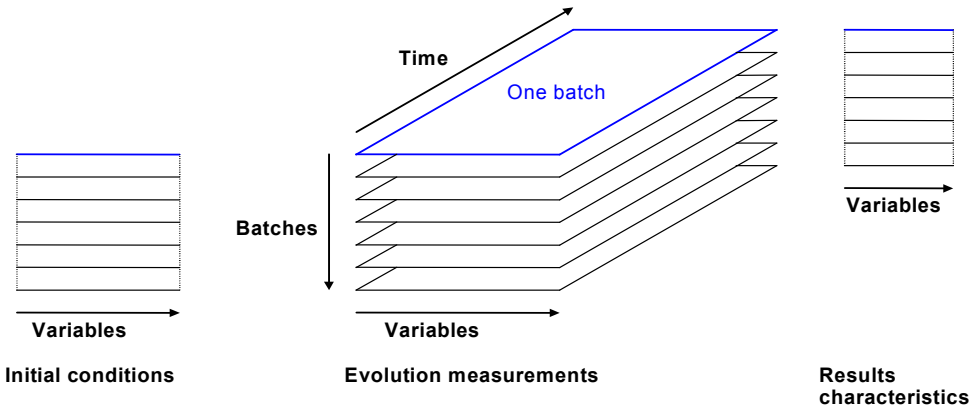


Figure 16.3: Batch-data often involves three distinct blocks of data, i.e., initial conditions (the Z-matrix), evolution measurements (the X-matrix), and results and quality characteristics (the Y-matrix). These data tables can be analyzed independently with PCA or be related to each other by PLS or OPLS.

16.2.3 Multivariate analysis is crucial in batch monitoring

Obviously, from a data analytical perspective, batch processes are quite different from continuous processes because of their three-way nature. However, as with continuous processes, the on-line computers that are connected to most batch processes routinely collect massive amounts of data. This is done for variables like pressures, temperatures, reactor tank levels, and flow rates. Hence, multivariate data analytical approaches are as crucial in batch monitoring as they are in the monitoring and diagnosis of continuous processes [Wold, *et al.*, 1998b; Nomikos and MacGregor, 1995; Dahl, *et al.*, 1999; Louwerse, *et al.*, 1999; Singhal and Seborg, 2002; Undey and Cinar, 2002; Kourti, 2003; Kourti, 2005; Aguado, *et al.*, 2007; Duchesne and MacGregor, 2009; Garcia-Munoz, *et al.*, 2010; Andersen and Runger, 2011].

Multivariate analysis of batch data enables the correlation structure among measured variables to be explored. This is important not only for gaining an understanding of the underlying properties that dominate a batch process, but also for early fault diagnosis. However, not only is the relationship among all batch variables at any time of prime importance, so also are the trajectories of all these variables. The development of various trajectories of batches may often serve as a fingerprint of each batch. Thus, by multivariately modeling average batch trajectory features, and deviations from normal batch evolution, it is often possible to separate good batches from bad batches.

Moreover, in addition to diagnosing whether a batch will end up as accepted or rejected, it is of great significance in batch monitoring to predict maturity of batches. With the word *maturity* we mean the extent to which a batch life-cycle has been completed. In simple cases, the state of a batch is an uncomplicated linear or non-linear function of time. However, in most cases batches develop differently due to varying influences of uncontrollable external factors. It is then sensible to use a status variable in order to monitor batch maturity. A batch process may also include several phases, i.e., localized time segments where kinetics and physico-chemical mechanisms of batches may vary considerably. This means that multivariate data analytical tools must be used, which allow separate multivariate models of the different phases, if needed.

16.2.4 Objectives of multivariate batch modeling

As discussed above, multivariate analysis enables an overview of all process data in one place. This overview includes the variability in all relevant measured process variables. One type of overview concerns within batch variations during processing and is accomplished with a so called batch evolution model (BEM). The other type of overview concerns between batch comparison after batch completion and is accomplished with a so called batch level model (BSM). A more detailed description of the BEM and the BSM concepts is given in Section 16.3.2.

Within multivariate batch modeling there are two typical cases, i.e., using developed models for (i) *process monitoring* or (ii) *process development*. The modeling strategies differ depending on the objective. When process monitoring is the target, the main objective is to ensure that the production is stable by identifying, interpreting and possibly correcting abnormal behavior in the process. When process development is the target, the main objective is to acquire better process understanding by identifying the influential process variables and any recurring upsets. Part of the work here may also deal with issues like comparison of scales, campaigns, sites, etc., to pinpoint differences and ensure similarity.

The use of multivariate batch models for *process monitoring* and *process development* have some commonalities as far as the multivariate modeling workflow is concerned, but there are also distinct differences. In the next few sections, we will review and exemplify the multivariate approach to BSPC. We will explain the batch evolution model (BEM) and the batch level model (BLM) in detail and point out how to apply such models in process monitoring and process development.

16.3 A multivariate approach to BSPC

In the following sections, we will introduce a PCA, PLS and OPLS-based strategy for batch modeling [Wold, *et al.*, 1998b]. Each consecutive step of the strategy will be exemplified using the BAKER'S YEAST example.

16.3.1 Example dataset: BAKER'S YEAST

This dataset concerns the production of baker's yeast, and was kindly contributed by Jästbolaget AB in Solna, Sweden. The manufacturing of baker's yeast is carried out in five steps. The duration of the last step is 14 hours, and this is the step with which we will be concerned. The dataset contains data for $N = 33$ batches, of which 16 were selected as reference batches. There was a great deal of variability among the batches arising from the molasses used, temperature, pH, etc. Seven variables ($K = 7$) were monitored and readings were taken every tenth minute, giving a total of $J = 83$ time points. The measured variables were: ethanol content, temperature, feed of molasses, feed of NH_3, air flow, level in tank, and pH.

Process monitoring of this batch process was deemed difficult due to fluctuations of the variables over time, but also because their mutual relationships were altered during the progress of a batch. Two examples are shown in Figures 16.4 – 16.5. As shown by these figures, there is considerable variation within each batch, and between batches. The main objective of the multivariate process study was, therefore, to examine if fault assessment could be accomplished at an early process stage, and preferably much earlier than could be achieved using a univariate approach.

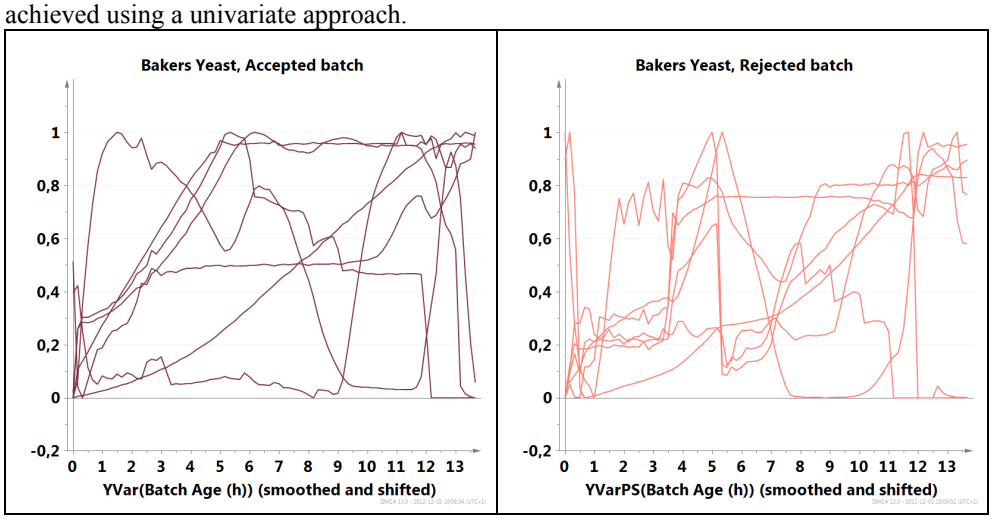


Figure 16.4: (left) Trajectories of seven variables for an accepted batch of yeast. Figure 16.5: (right) Time series plot of seven variables for a rejected yeast batch.

16.3.2 Two types of batch models

The present approach is based on two types of batch models, the *batch evolution model (BEM)* and the *batch summary model (BSM)*. With the batch evolution model, the interest lies in (i) evaluating the individual observations, i.e., time points, (ii) predicting batch maturity, and (iii) understanding the typical evolution trace of a normal, good batch. Thus, the aim is to develop a model of the development of good batches, use it to monitor new batches as they evolve, and ideally accomplish early fault detection. With the batch level model, all available data, i.e., initial conditions (Z), batch process variables (X), and result and quality characteristics (Y), are used for making a model over the whole batch. This whole batch model is used after batch completion to understand between batch variability. Another important objective is to understand how Y is influenced by the combination of initial conditions and batch evolution.

16.4 The batch evolution model (EXAMPLE: BAKER'S YEAST)

16.4.1 Unfolding of data preserving the variable direction

In order to accomplish batch evolution modeling, the three-way batch data table must be unfolded. Unfolding means that a higher-order data structure is cut into slices, which, in turn, are ordered sequentially according to a lower-order data structure. The present approach is based on unfolding the three-way training set data matrix such that the direction (the mode) of the variables is preserved (Figure 16.6). The resulting two-way matrix then has *N* x *J* rows (observations) and *K* columns (variables). Hence, in this matrix, X, the units are the individual observations, not the entire batches. In the BAKER'S YEAST dataset there are altogether 33 (batches) times 83 time points = 2739 observations. These observations are mapped by seven variables. A PCA on this X-matrix will show how the individual observations relate to each other.

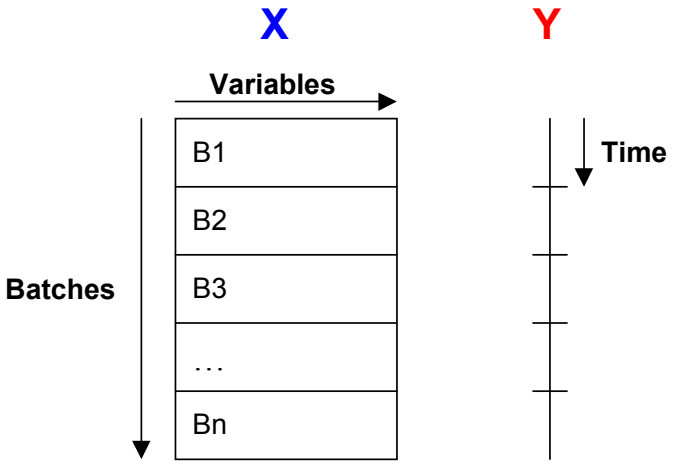


Figure 16.6: The three-way data table of Figure 16.2 is unfolded by preserving the direction (the mode) of the variables. This gives a two-way matrix with N x J rows and K columns. Each row contains data points x_{ijk} from a single batch observation (batch i, time j, variable k). If regression is made against local batch time, the resulting PLS scores reflect linear (t_1), quadratic (t_2), and cubic (t_3) relationships to local batch time.

Furthermore, as indicated by Figure 16.6, it is possible to include batch age or batch time in the modeling. In the Baker´s Yeast example there is a variable denoted Batch Age, and that variable will be used as the y-variable in the PLS model discussed below. Unless supplied externally as a real time-variable, the SIMCA software has the capability during data import to auto-generate a dummy y-variable reflecting relative local batch time. This dummy variable is called $Time. The value of $Time predicted by the PLS model is a very suitable maturity index that may be used to indicate how far a batch has evolved, if it is ready for termination, and so on. The resulting PLS scores are new variables that capture linear (t_1), quadratic (t_2), and cubic (t_3) relationships to local batch time. Alternatively, one may choose to include a batch maturity index as a y-variable, if such information is present. When PLS

is used to relate X and Y, the objectives are to predict the state of the process, and to understand batch phase.

16.4.2 Creating a BEM of the 16 accepted batches

The 16 accepted batches representing good past batch process behavior were used to build a BEM of X versus Batch Age. Cross-validation indicated the significance of three PLS factors, corresponding to $R^2X = 0.85$, $R^2Y = 0.97$, and $Q^2Y = 0.97$. Figure 16.7 displays the PLS t_1/t_2 score plot, where the observations pertaining to one batch have been connected by a line. Apparently, there is strong behavioral similarity among the different batches. In general, each batch starts in the lower left-hand quadrant of the score plot, and ends in the right-hand middle part. At first, there is a steady upward trend for each batch along the t_2-direction. Generally, this trend continues until batch age 6 hours, whereupon a sharp shift in process evolution in a "southeasterly" direction takes place. An interpretation of this model is possible by looking at the corresponding PLS weights plotted in Figure 16.8. We can see that "level in reactor tank" is the X-variable most strongly correlated with batch age. The variables "level in tank", air flow, pH, and temperature change in the horizontal direction. This tells us that the numerical values of these four variables increase during batch evolution. In the vertical direction, ethanol content and air flow dominate. These two variables undergo large changes when the shift in "process evolution direction" begins.

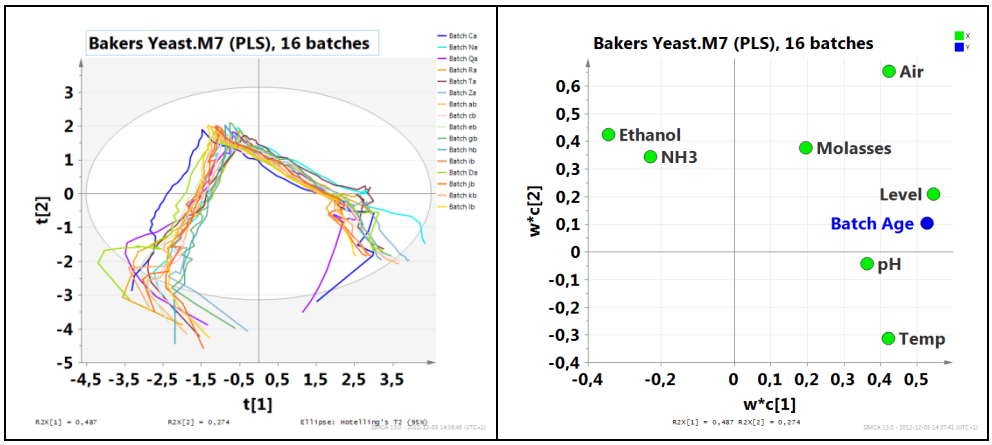


Figure 16.7: (left) PLS t_1/t_2 score plot showing the evolution of the 16 reference batches of the BAKER'S YEAST dataset. Scores corresponding to batch initialization are found in the lower left-hand part. Variation in batch termination can be seen in the right-hand part. Figure 16.8: (right) PLS $w*c_1/w*c_2$ weight plot corresponding to Figure 16.7. Batch age is positively correlated with level in tank, air flow, pH, and temperature. The response variable is only slightly correlated with feed of molasses, ethanol content, and feed of NH_3.

16.4.3 Rearrangement of scores to give batch control charts

In order to establish control charts for monitoring the development of new batches, some re-arrangement of BEM scores is necessary. Firstly, the three score vectors (three in the case of BAKER'S YEAST) are divided in segments corresponding to each batch in the training set data. Secondly, these segmented score vectors are re-ordered such that the scores of one batch form one row vector of a new score matrix X_T, i.e., a new matrix where, in each row, t_1 of one batch is followed by t_2 of the same batch, which is followed by t_3 of the same batch, and so on. This new matrix X_T contains as many rows as there are batches in the training set data, i.e., 16 in the BAKER'S YEAST example.

By computing the average and standard deviation of each series of score vectors (see Figure 16.9), it is possible to derive control limits for batch evolution of good batches. It seems appropriate to set control limits at average score ±3 SD at each time point, as is normally done in MSPC. The essence of this re-ordering principle is that we have obtained, for each component of the initial PLS model, an average trajectory with upper and lower control limits. These statistics indicate the normal developmental traces of new batches.

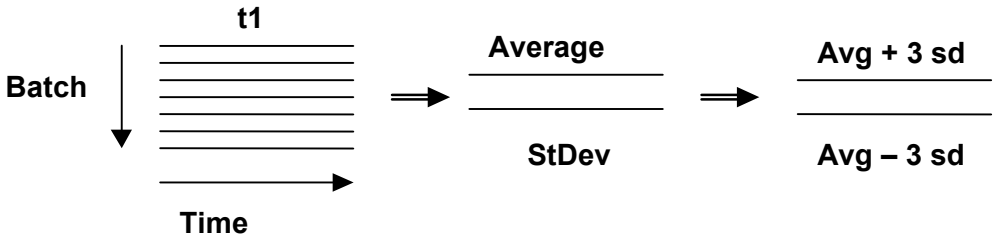


Figure 16.9: The score values for each batch are arranged as row vectors underneath each other, giving a new matrix X_T that has the number of rows equal to the number of batches in the reference dataset. Observe that the drawing only outlines the re-arranging principle for the first score t_1. From this new matrix one calculates the averages and standard deviations (SDs) of the matrix columns, and subsequently control limits as averages ± 3 SD.

Not only can the PLS model scores be charted as described above, so also can e.g. DModX and predicted values of batch age or batch time. DModX shows the unmodeled variation, and a good new batch should evolve in the same way as the reference batches and be inside D_{crit}. In addition, the predicted value of batch age or batch time can be used as a status index to indicate current batch maturity of new batches.

16.4.4 Batch control charts of the BEM

The resulting batch control charts of the BAKER'S YEAST scores are shown in Figure 16.10. When examining these scores we should remember their explanatory power for the seven measured variables. In turn and order, the three scores t_1-t_3 account for 49%, 27%, and 9% of the total variation. Hence, it is important that no anomalies are seen in the control charts of the first two scores. Milder, temporary deviations outside control limits in the higher-order scores would not necessarily cause alarm.

According to Figure 16.10, in the first batch control chart (t_1) there is a steady transition from low scores at batch initialization to high scores at batch termination. All reference batches behave well according to the upper and lower tolerance limits. We already know from the previous model interpretation that the first score well reflects batch age. Moreover, in the second batch control chart (t_2), we can see the change in batch mechanism occurring at batch age 6 hours. However, all reference batches are inside the control limits. All batches are also mostly inside the control limits in the third batch control chart (t_3).

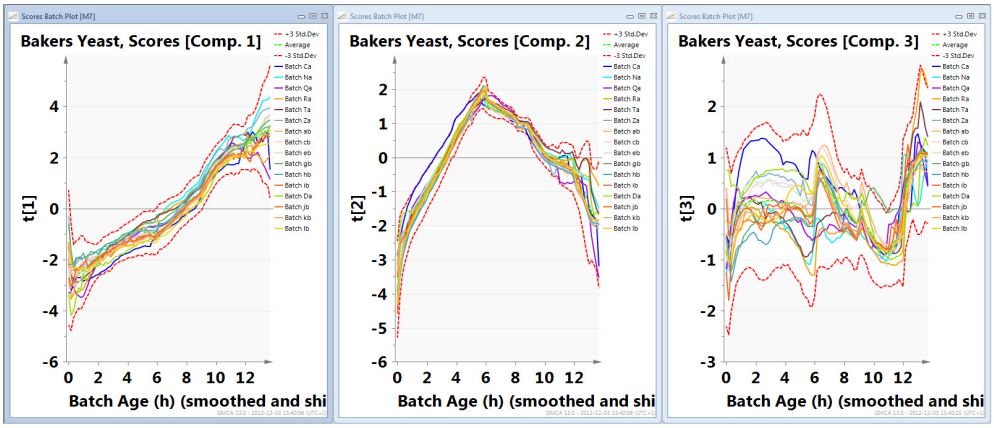


Figure 16.10: Control charts of the three PLS t-scores for the BEM of the BAKER'S YEAST dataset. In turn and order, the three scores t_1-t_3 account for 49%, 27%, and 9% of the total variation in the seven batch process variables. The conclusion from these three control charts is that, among the 16 selected reference batches, no batch displays conspicuously odd behavior during their maturation.

The two control charts plotted in Figure 16.11 refer to predicted batch age and DModX. If the distance to the model is too large, the correlation structure has changed. This is commonly the most sensitive indicator of process upsets. As shown by the DModX control chart, none of the 16 reference batches persistently stays outside the control limit. Also, it is interesting to see that predictions of batch age well conform with "observed" batch time up

to approximately 12 hours. In general, if the predicted time (maturity) is higher/lower than the actual, the batch is progressing too fast/slow. The predicted batch age at late-stage batch evolution shows substantial variation compared with early-stage estimates. This means that the seven batch process variables show much more varied trajectories in late-stage batch development than in the initial stages.

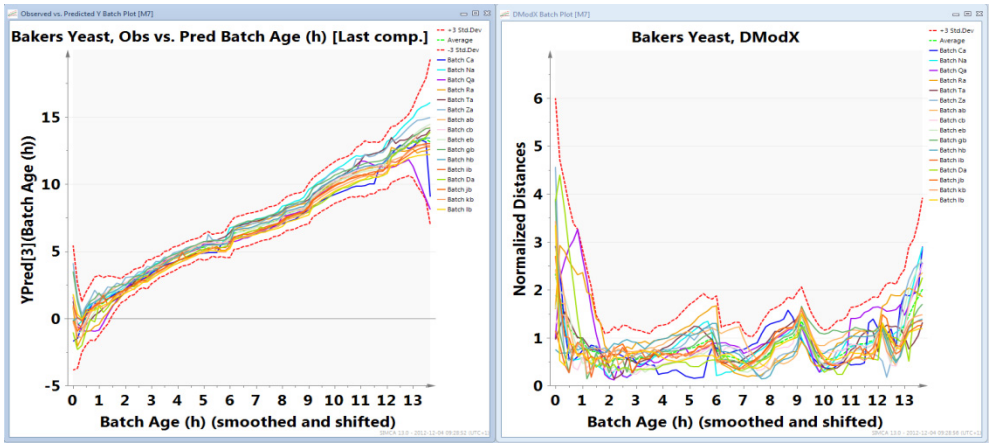


Figure 16.11: Batch control charts of predicted batch age (right) and DModX (left). The predicted batch age can be used to assess progress and maturity of new batches as they evolve. The DModX control chart can be used to assess whether variable correlation patterns of new batches conform with or deviate from the pattern manifested by the 16 reference batches.

16.4.5 Monitoring the evolution of new batches

The established control charts can be used to monitor the development of new batches. Any observed deviation from normal traces can be interpreted by means of contribution plots. Here, we will show results for complete batches. However, the principle is the same for batches which are still underway. Batches still in progress may be followed in batch control charts, even though all data have not yet been determined.

Although the BAKER'S YEAST dataset contains data for 17 test set batches, in this section we will only display monitoring results for three test set batches (Figure 16.12). This is done for reasons of clarity in the plots. The three batches that we have selected have markedly different evolutionary profiles. Below, these batches are referred to as batch Ba, Ja, and Pa.

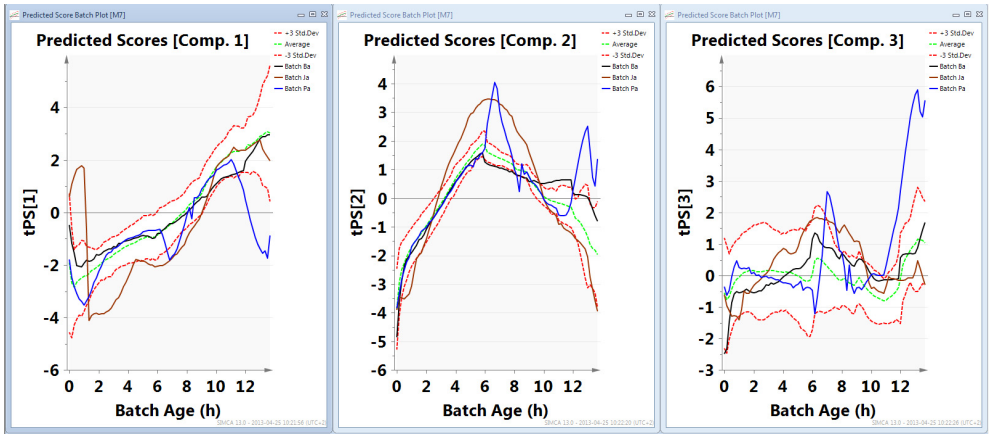


Figure 16.12: Batch prediction control charts of scores t_1 – t_3. Predicted trajectories of the three test set batches (batch Ba, Ja, and Pa) are given.

As demonstrated by Figure 16.12, one of the batches (batch Ba) mostly develops correctly and is found inside the upper and lower control limits in all three batch control charts. Another batch (batch Ja) experiences problems in the initial and middle parts of batch evolution, but recovers and is OK at the end. We can particularly see the deviation of batch Ja in the control charts of t_1 and t_2. The last predicted batch (batch Pa) deviates toward the

late stages of production, where it produces ethanol instead of baker's yeast. This deviation from the pattern given by the reference batches is pronounced for all scores (t_1, t_2, and t_3).

In addition to batch control charts of scores, an informative graph is created when DModX is charted against time (Figure 16.13). Actually, in the BAKER'S YEAST example, the most clear-cut indications of batch deviation are found in the DModX control chart. According to Figure 16.13, one batch diverges substantially from the others at an early stage. This is batch Ja. Another batch (batch Ba) develops inside the control limit almost all of the time, whereas the last test set batch (batch Pa) drifts away in late-stage production.

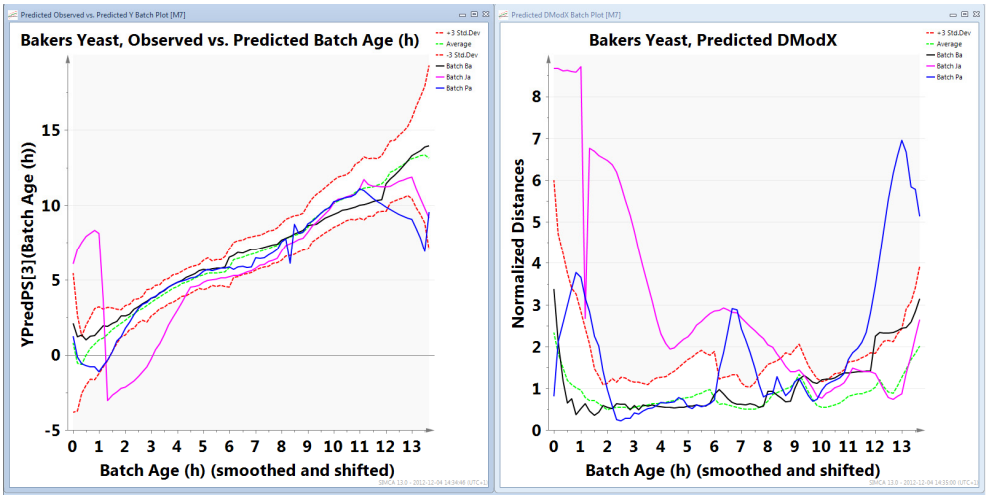


Figure 16.13: Batch control charts of predicted batch age (left) and DModX (right). Except for a slight deviation of batch Ja at the beginning, all batches behave well according to the maturity control chart. The DModX control chart clearly shows how batch Ja exceeds the DModX control limit at low time points, and how batch Pa exceeds the control limit at high values of local batch time. The last test set batch (batch Ba) is located inside the control limits throughout the progress of the batch, except for a short deviation occurring at 12 hours.

16.4.6 Finding variables contributing to deviation from batch normality

As indicated by the DModX control chart in Figure 16.13, batch Ja was outside the control limit in the initial batch development, and batch Pa towards batch completion. Similar information was obtained from the batch control charts of scores rendered in Figure 16.12. In order to understand which variables are probably responsible for these deviations, we may use contribution plots. We say "probably" because the only way to be absolutely sure and obtain conclusive evidence (i.e., a causal relationship) is through the use of design of experiments (DOE).

Figures 16.14 and 16.15 provide two contribution plots (scores mode). The first contribution plot relates to early batch environment and shows, at batch age 1.33 hours, how batch Ja and the average batch differ (Figure 16.14). Evidently, the large difference resides in just one of the seven variables, namely pH. This variable is far lower (i.e., conditions are more acidic) for batch Ja than for the average batch. After bringing pH under control, batch Ja recuperates and ends up in satisfactory condition (cf. Figures 16.14 and 16.15).

The second contribution plot relates to differences between batch Pa and the average batch at batch age 12.17 hours (Figure 16.15). It does not take long to realize that ethanol content is very different in batch Pa. In fact, the ethanol content of batch Pa is almost five standard deviations higher than that of the average batch. It is known that batch Pa failed because it produced ethanol rather than yeast, so the information extracted from the presented contribution plot makes sense.

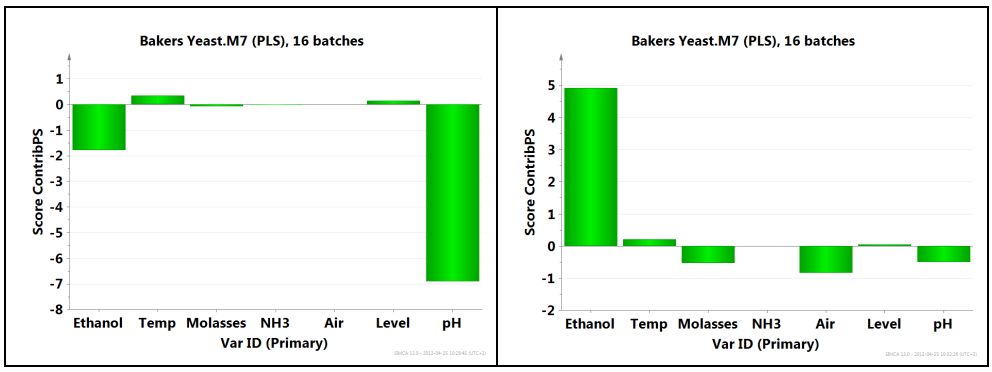


Figure 16.14: (left) Batch contribution plot (scores mode) of how the test set batch Ja and the average batch differ at batch age 1.33 hours. The variable pH is lower for batch Ja shortly after batch commencement. Figure 16.15: (right) Batch contribution plot (scores mode) of how the test set batch Pa and the average batch differ at batch age 12.17 hours. The ethanol content is much too high for batch Pa.

16.4.7 Summary of batch evolution modeling

In Section 16.2.4 we discussed the objectives of multivariate batch modeling and the two typical cases of process monitoring and process development. It is important to point out that a batch evolution model (BEM) can be used in both cases. In process monitoring the BEM should focus on what is known to be good and therefore it will embrace limited process variability. Thus, only well performing and well behaving batches are selected as the basis for the BEM. By contrast, in process development a much broader spectrum of batches is selected as the basis for the BEM. This is because process development is a more multifaceted task and may correspond to several different situations, such as, understanding drifts in a process, investigating differences between sites or scales, and characterizing raw material impact. For each situation, it is imperative that the BEM must contain all sources of relevant information and often more batches are needed than in process monitoring.

In process monitoring, the main aim of the batch evolution level modeling is to obtain a model of the development of good batches, use it to monitor new batches as they evolve, and ideally accomplish early fault detection. In this type of modeling, three-way batch data are unfolded to preserve the variable direction (Figure 16.6). This means that each row of the resulting two-way X-matrix corresponds to one time point of one unique batch. PLS is used to relate this process information to a y-variable representing batch time or batch maturity. One then obtains a maturity index model that is used to explore how a batch has progressed. The resulting PLS X-scores (t_1, t_2, …) summarize variables with a dependence on batch time or maturity.

By re-arranging the resulting PLS X-score vectors and other model statistics (DModX, predicted time or maturity) according to the scheme displayed in Figure 16.9, it becomes possible to create batch control charts. Observe that no new PLS model is calculated; we simply average the already computed model parameters column-wise and compute the standard deviation around each average. The resulting control charts can be used to follow the trace of a developing batch, and extract warnings when it tends to depart from the typical trace of a normal batch.

16.5 The batch level model (EXAMPLE: BAKER'S YEAST)

16.5.1 Unfolding of data preserving batch direction

In Section 16.5 we will address the batch level model (BLM). In this model all available data, i.e., initial conditions (*Z*), batch process variables (*X*), and result and quality characteristics (*Y*), are used for making a model over the whole batch (Figure 16.16). Depending on the objective and available data the BLM can either be a PCA model or a PLS model. In the current example we will exemplify a PCA BLM, in the next example (the Wafer dataset) we will exemplify a PLS BLM.

In contrast to the BEM data table, where each row corresponded to one time point in one particular batch, each row in the BLM data table represents one whole batch. To account for the batch evolution properties it is recommended to use the original process variables; however, it should be noted that it is possible to use the scores of the preceding BEM. If the original process variables are used, the three-way batch process data table is actually unfolded to preserve the batch direction.

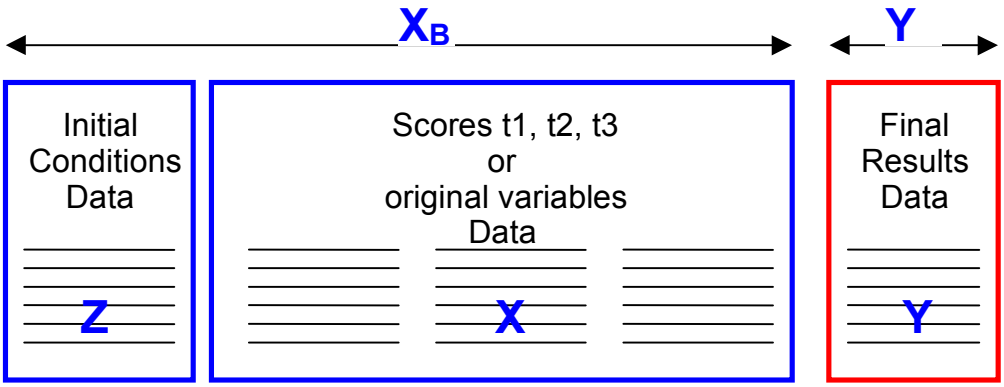


Figure 16.16: In the batch level model, all available data are used to obtain a model of whole batches. Note that each row corresponds to one batch. Initial conditions data are often pooled with process evolution data to form a new X-matrix, X_B, which can be summarized using PCA. Provided that Y-data are available, this X_B-matrix may also be related to the final results using PLS.

16.5.2 Computing a BLM on the 16 reference batches

In the BAKER'S YEAST dataset there is an additional variable relating to initial conditions, i.e., Innoc, which represents the total amount of dry substance added at start (in kg). We will here use PCA to summarize the information in the initial condition (Innoc) variable and the process information, i.e., the seven rearranged process variables. The resulting four-dimensional BLM explained 72% and predicted 32% of the process variation.

Figures 16.17 and 16.18 show scatter plots of the four PCA score vectors. In both plots, the 16 points are rather evenly scattered, which indicates a homogenous group of reference batches. All batches are inside the tolerance ellipse. The corresponding DModX statistic does not indicate any deviating batches (no plot shown).

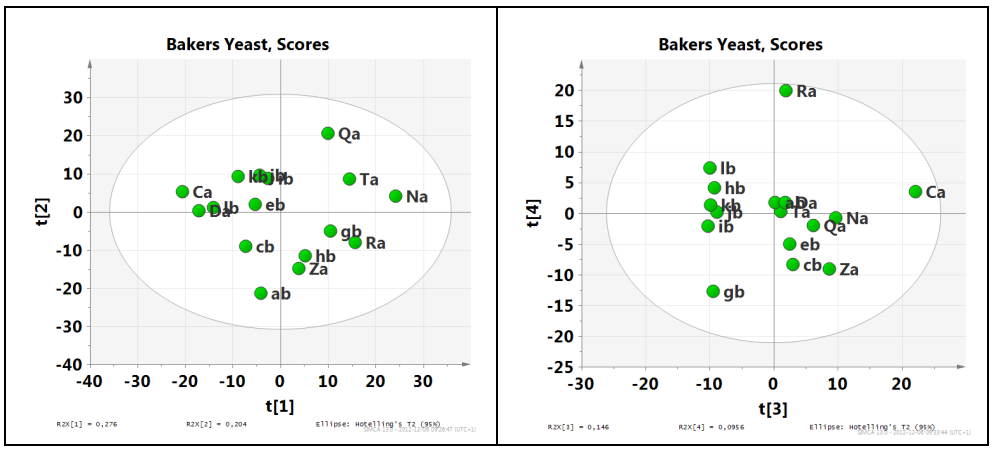


Figure 16.17: (left) BLM t_1/t_2 score plot. Each observation (batch) is represented by one point. The scatter plot shows a homogenous distribution of the batches. Figure 16.18: (right) BLM t_3/t_4 score plot.

16.5.3 Model interpretation

One way of interpreting which X-variables are most relevant in the BLM, consists of looking at the sources of variation plot for the first principal component (Figure 16.19). In

this type of plot, the changing influence of a process variable across the batch lifetime is comprehensively documented. Similar plots can of course be created for the remaining three principal components, but since the first component is the most important one, we focus the interpretation on this component. In general, the strongest contribution to the BLM is displayed by the process variables level in tank, ethanol content, feed of molasses and feed of NH_3. Some process variables are influential during shorter time segments of batch lifetime whereas others stay influential throughout the entire time span.

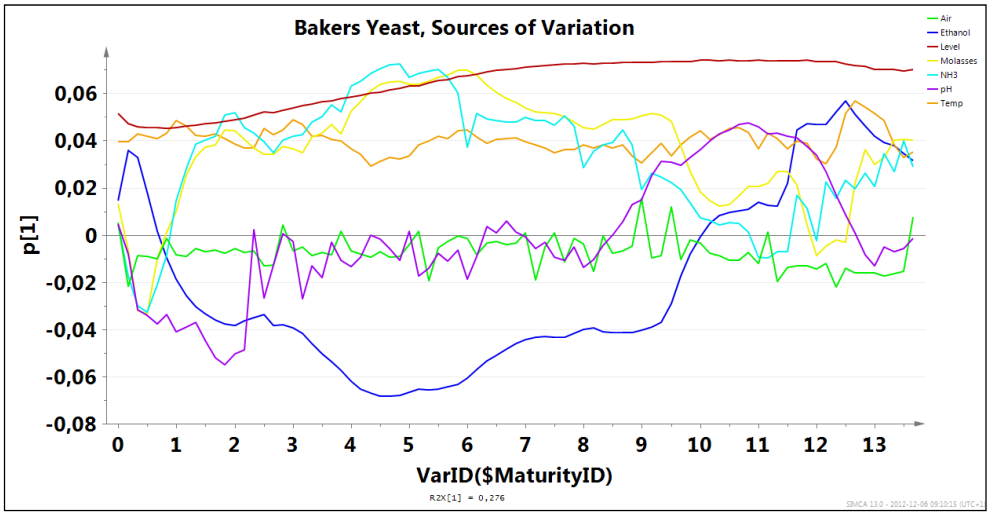


Figure 16.19: Sources of variation plot for the first component of the BLM. This plot is color coded according to process variables. Similar plots can be created for the remaining three principal components.

16.5.4 Classification of new batches

The established BLM can be used to classify new batches. In total, there are 17 batches in the test set, some of which gave good product but also some that failed for various reasons. We created the predicted DModX plot seen in Figure 16.20 to overview the classification. The plot is color coded according to whether the batch belongs to the training set or the prediction set. The large majority of test batches are positioned quite far above the Dcrit limit of the DModX plot, indicating their difference to the 16 reference batches. Additionally, the most deviating batch (Ya) is highlighted to illustrate how contribution plotting can be used to interpret why a batch deviates from normal behavior. Figure 16.21 shows the contribution plot for Ya. Apparently, ethanol content is much larger for this batch than for the 16 reference batches.

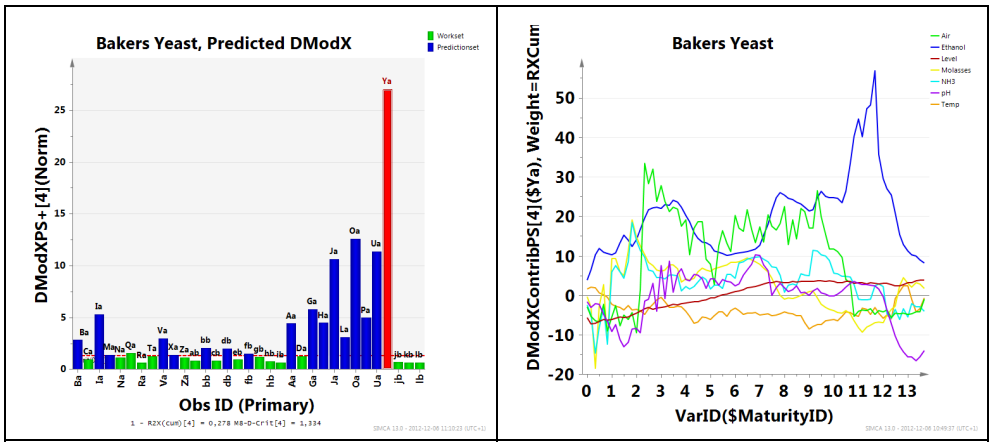


Figure 16.20: (left) Column plot of predicted DModX. Green color represents the training set. Blue color represents the prediction set. One highlighted batch, batch Ya, is show in red. Figure 16.21: (right) Contribution plot, DModX-mode, suggesting which variables cause batch Ya to deviate substantially from normal batches.

16.5.5 Summary of batch level modeling

As has been pointed out in Section 16.5, the batch level modeling in BSPC involves using all available data to make a whole batch level model (BLM), i.e., to use initial conditions (Z), batch evolution data (X), and result and quality characteristics (Y). Depending on the available data and the goal of the model, this BLM can either be a PCA or a PLS model. The BLM is based either on the scores of one or more BEMs, or the original variables, and can be used to classify prediction set batches as good or bad. Another important objective is to understand how Y is influenced by the combination of initial conditions and batch evolution.

Here, 16 reference batches were used to train a model to recognize "good" operating conditions. This model was able to categorize between good and problematic batches. Moreover, an interesting modification of the presented approach is conceivable. At regular time intervals, e.g., at 10%, at 20%, etc., of the batch evolution, the resulting BEM score vectors can be re-arranged and used to form a "partial batch level model". This model could then be used to infer whether the evolution pattern of the new batch is consistent with normal batch trajectories, and also to obtain very preliminary estimates of the final results and quality responses (Y).

16.6 A second batch example: Wafers

We will now illustrate the proposed approach to BSPC by a second example, called Wafers.

16.6.1 Background and arrangement of data

This example is taken from the Wafers dataset used in Chapter 11. Recall that the dataset is from an oxide chemical vapor deposition (CVD) process and that in Chapter 11 the data were analyzed using the average values from each phase. In this section we will use all of the process time points collected during the processing of the batch (wafer). The full dataset is a three-way data table with the "modes" batches (N) * time (J) * variables (K). K = 13 variables were measured on N = 2148 wafers during the 12 steps (phases) of the tool. The 13 variables were measured on-line every second and the phases range from a couple of seconds to 50 seconds long. After the completion of each wafer, one quality variable, oxide thickness, was measured. The variables include gas flows, temperatures, and pressures all measured during every phase. Five of the phases were steady state for most variables and seven were transients.

Prior to the data analysis two entire phases were deleted because there were less than three observations per batch for those phases. The length of each phase is different between batches, but doesn't vary by more than 10%, so we used the time variable, $Time, generated by the SIMCA software as maturity index. A set of 1067 wafers was randomly selected to represent normal process behavior. The remaining 1081 wafers were selected as a test set.

16.6.2 Objectives of the BSPC study

Using the Wafers data, our first objective will be to develop a model of the evolution of the 1067 batches (the batch evolution model, BEM), and use it to monitor new batches, as they evolve, in order to detect problems as early as possible. This is called early fault detection. Our second objective will be to construct a batch level model (BLM) of the whole batch duration based on the BEM, and use this second model to predict oxide thickness.

16.6.3 Batch evolution model

The three-way data table was unfolded to a two-way data table. Ten PLS models, one for each phase, were fitted between the measured variables and time (the auto-generated $Time) for the 1067 training set batches. These models were very efficient in modeling batch behavior and explained X-variances were close to or exceeded 70% for all phases. The scores of the BEM (consisting of the 10 phase specific PLS models) were used to create batch control charts, using the basis shown in Figure 16.9.

In SIMCA we can create a series of different control charts, i.e., charts relating to score vectors, DModX, predicted time, etc., all highlighting different aspects of how the batches evolve. Figure 16.22 displays control charts of the first score for two of the phases. All batches in the training set are within or near the control limits in these score control charts. Remember that the control charts shown here display how the individual batches evolve.

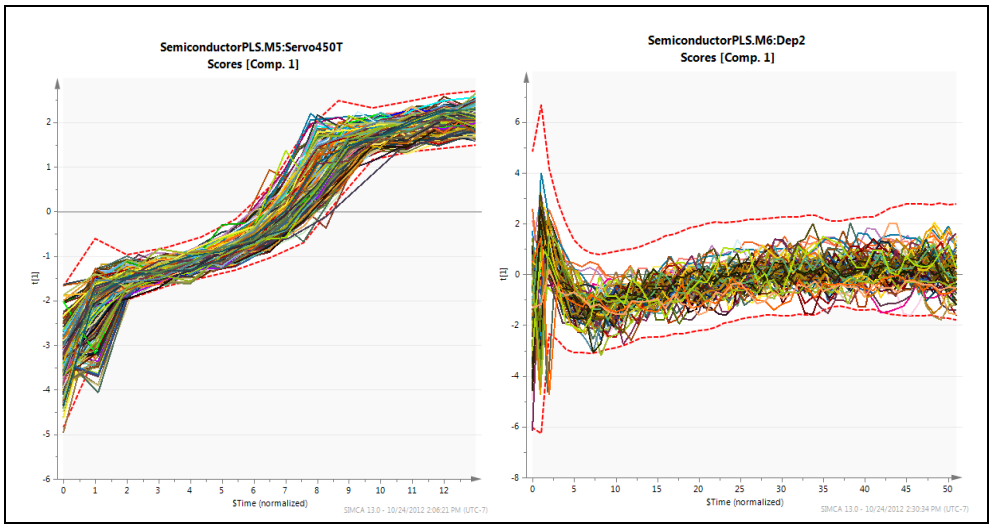


Figure 16.22: Batch score control charts of t_1 for two phases of the Wafers dataset. All reference batches are near the control limits. Hence, there is no indication of extreme batches among the 1067 chosen for model training.

16.6.4 Use of BEM: Classification of test set batches

The next step is to utilize the batch evolution model and test its ability to discriminate between good and bad test set batches. This will give us an indication of how the model will perform in real-time to detect process deviations. For reasons of brevity, only a small selection out of many possible control charts will be plotted.

We will start with the good batches. Figure 16.23 displays the two predicted scores batch control charts, i.e., control charts with the same control limits as in Figure 16.22 but used to follow the paths of 300 new batches. The 300 test set batches are mostly inside the control limits in both score control charts. They are also well-behaved according to DModX and predicted local batch time, but such plots are not shown. Hence, we may infer that the batch evolution model can classify accepted batches as good.

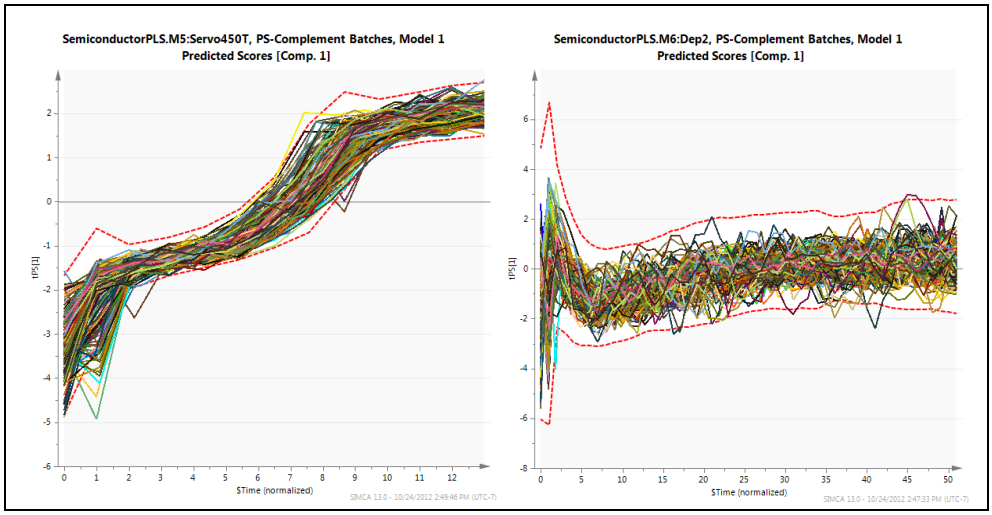


Figure 16.23: Batch predictions control charts of scores t_1 for two phases, employed for evaluating the trajectories of 300 good batches. The batches are inside or close to the control limits (red dashed lines), and are therefore classified as good.

Next, we move on to some problematic batches. Figure 16.24 re-displays the t_1 batch control chart showing the evolution of some outlying batches for the main process step, Dep2. This plot demonstrates the ability of the model to identify deviating batches.

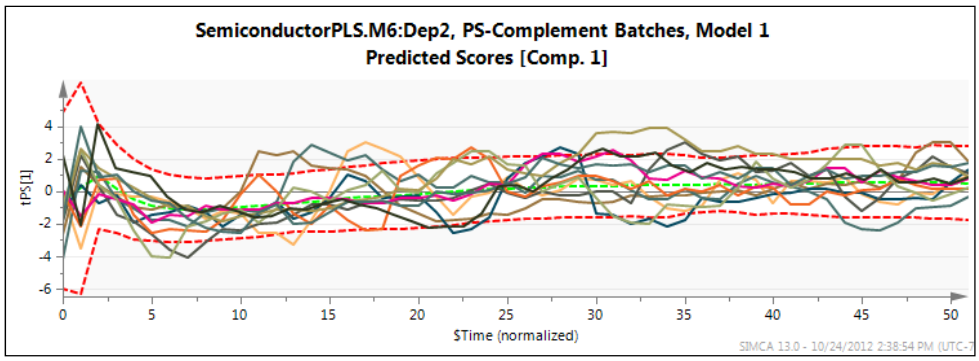


Figure 16.24. Batch predictions control chart of score t_1 giving the trajectories of some bad batches. These batches have more variation in scores and are outside the control limits on several occasions and are therefore classified as bad.

Contribution plotting may be used to interrogate the model and highlight which underlying variables are actually contributing to an observed process upset. As an example, Figure 16.25 uncovers that the ThrottleValve variable is too low at the early stage of the Dep2 phase for one of the wafers. As reference, we chose the average batch. Again, we stress that what we see in these plots are variables in which certain batches deviate numerically. Remember that causal relationships can only be established by design of experiments (DOE).

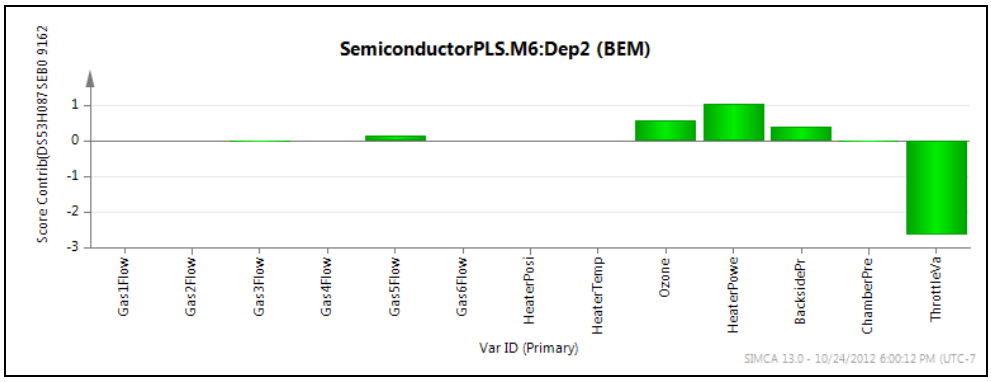


Figure 16.25: A contribution plot suggesting that variable ThrottleValve is too low in the Dep2 phase for this batch.

16.6.5 Batch level model

To accomplish a model of the evolution of the whole batch, the score vectors of the ten concatenated lower level PLS models were rearranged as depicted in Figure 16.16 where each row corresponds to one batch. Because there is an external y-variable available, oxide thickness, we will build a PLS batch level model to do the prediction. The resulting PLS model contains four significant components utilizing 44% of the process variation to describe 82% of the variance in oxide thickness. This means that there is enough correlation between the process data and the thickness measurement to build a prediction model. Figure 16.26 shows a scatter plot of t_1 versus t_2. There is some obvious clustering by chamber which is usually the case with semiconductor data. There are also some minor outliers which should be investigated for process understanding. This exercise is not done here. Figure 16.27 displays the relationship between observed oxide thickness and calculated (by PLS model) oxide thickness. The values of the thicknesses are hidden due to confidentiality. Again, there is clustering by process chamber which, although not ideal, did not prevent the model from making predictions within an acceptable range.

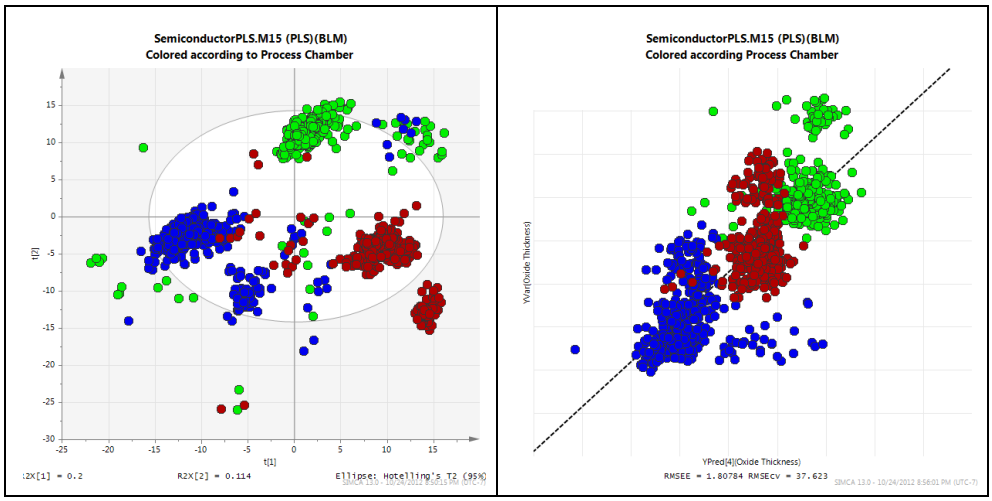


Figure 16.26: (left) PCA score plot of t_1 and t_2 of batch level model of Wafers dataset. Each observation (wafer) is represented by one point. Figure 16.27: (right) Relationship between measured and calculated value of oxide thickness.

Eleven of the variables appear to be important for the model at the batch level, with the pump step being the most important overall. Figure 16.28 provides the batch variable importance. This parameter is a function of the scores at the batch level and the PLS weights derived from the BEM. It shows the overall importance of each variable in explaining the quality (good/bad) of batches. We do not suggest using the batch variable importance to determine which variables to use in the model (or which to exclude). We do however recognize that knowing which variables have historically been predictive of quality can be useful information to the process engineer.

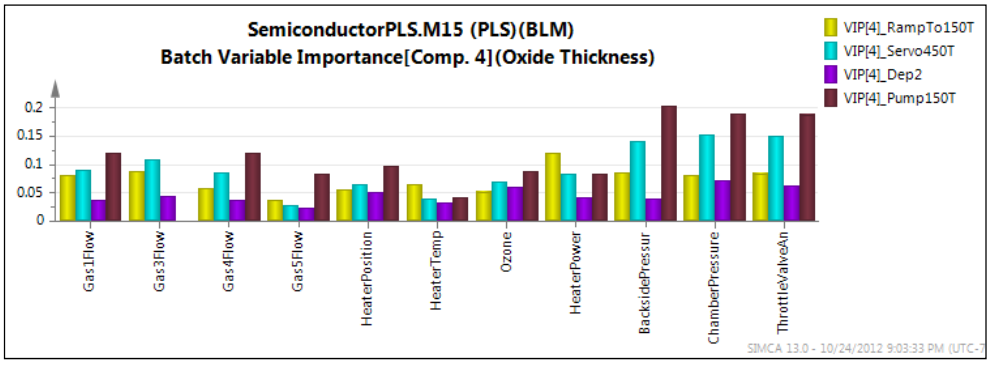


Figure 16.28: Plot of batch variable importance. This plot indicates the overall importance of each original variable in describing batch evolutionary properties.

16.6.6 Use of BLM: Validation using the test set

In order to validate the prediction model, the developed BLM can be used to predict oxide thickness for the test set batches. Many of the prediction wafers deviate from normality in a similar way to the workset according to the score plot (Figure 16.29). Figure 16.30 shows the observed versus predicted oxide thickness for the test wafers. The accuracy is very similar to that for the workset, so the model is validated.

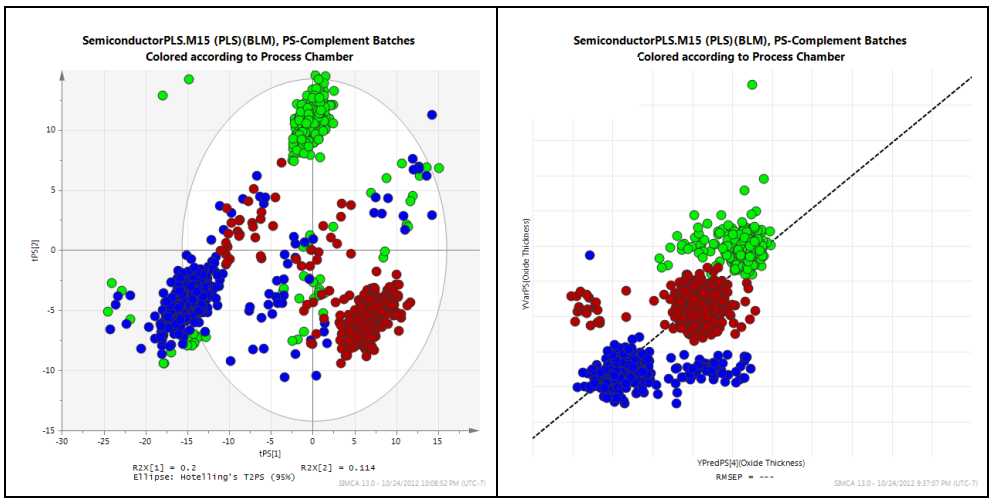


Figure 16.29: (left) Score plot of the batch level PLS model showing the location of the predicted test set batches.
Figure 16.30: (right) Observed versus Predicted oxide thickness for the test set batches.

16.6.7 Summary of Wafers application

In the Wafers example, the modeling of the evolution of a set of 1067 reference batches allowed us to construct control charts to monitor new batches during their evolution. The batch evolution model made it possible to see deviating batches. Probable cause for one of these process upsets was diagnosed by using contribution plots. A ThrottleValve variable was found to be too low at the early stage of the Dep2. Furthermore, the higher level model, the batch level model, permitted us to predict oxide thickness. This model was based on ten concatenated batch evolution models, each pertaining to one of the ten varying phases of the batch duration.

Hence, for this application, it is concluded that initial modeling of good reference batches makes it possible to derive useful batch evolution and batch level models which can discriminate between good and bad batches and predict process quality. The next step for this type of application is to test it online for a reasonable amount of time before enabling actions for alarms to further validate the model performance is as expected.

16.7 Questions for Chapter 16

1. What is a batch process?
2. What is BSPC? Why is it useful?
3. How are batch data often organized when they are received?
4. What is unfolding?
5. Which three types of batch data are discernible?
6. How are data organized in the BEM?
7. What are the objectives with the BEM?
8. Which plotting tool is useful for understanding how a batch deviates from normal behavior?
9. Which data can be utilized in a BLM?
10. What are the objectives with the BLM?

16.8 Summary and discussion

16.8.1 BSPC – How we do it

We have in this chapter described a multivariate approach to statistical analysis of batch process data. The main idea of this BSPC methodology is to use all available data for batch process modeling and monitoring. There are two levels, the lower one giving the batch evolution model, and the higher yielding the batch level model. The BEM means (i) working with individual observations, (ii) monitoring the evolution of new batches, and (iii) classifying current batch-phase. The BLM involves (i) working with whole batches, and (ii) predicting the outcome of evolving batches (good or bad). PCA, PLS and OPLS may be used at both levels.

For process monitoring, the models that are developed arise from a set of accepted batches representing good batch behavior and quality. These models provide a powerful means for overviewing and monitoring batches as well as making on-line predictions. Although many variables are dealt with in batch monitoring, the simplicity of presentation and interpretation of conventional SPC charts is retained in the new BSPC strategy. Deviating batch behavior can be diagnosed when charting multivariate parameters, like scores and DModX, and an interpretation of discovered process upsets is possible through, e.g., contribution plots. Dominating variables are found in a plot of batch variable importance.

16.8.2 Some advantages of the BSPC strategy

An advantage of this approach is that the batch evolution modeling can be carried out even if not all data have been measured, i.e., *even when a new batch is still underway*. This is in contrast to other multivariate methods for batch modeling, which usually assume that data for complete batches are available, and therefore must use some kind of missing data imputation of batches yet to be finished. Some related methods are discussed below (see Section 16.8.4).

Batch level modeling is conducted after the batch evolution modeling, usually using the data of the measured process factors. This is the recommended approach. Additionally, however, the multivariate BSPC approach also provides an alternative option to model the complete batch based on using a compressed version of the process factor, i.e. the scores. The use of the BEM scores in the BLM saves computer time and relieves data storage requirements, which is relevant in today's complicated batch process monitoring tasks where often very many variables are registered.

16.8.3 Going online

Batch models that are developed offline can be executed online using SIMCA-online. Figure 16.31 shows a screenshot stemming from the Baker´s Yeast application. Two example batch control charts are seen, that of the first score vector t_1 and that of DModX. The evolutionary profile of a new batch still under development is shown. As long as the new batch is inside the control limits of the control charts, the process operator knows that the batch is developing as it should. However, if there is a crossing of a control limit, the process operator receives an alarm drawing the attention to the problematic batch. By means of contribution plots and loading plots it is then possible for the process operator to drill down to the problematic variable or variables causing the process deviation. It should be observed that the SIMCA-online software can be used for both batch and continuous processes. A good example of an online application is found in [Streefland, *et al.*, 2009].

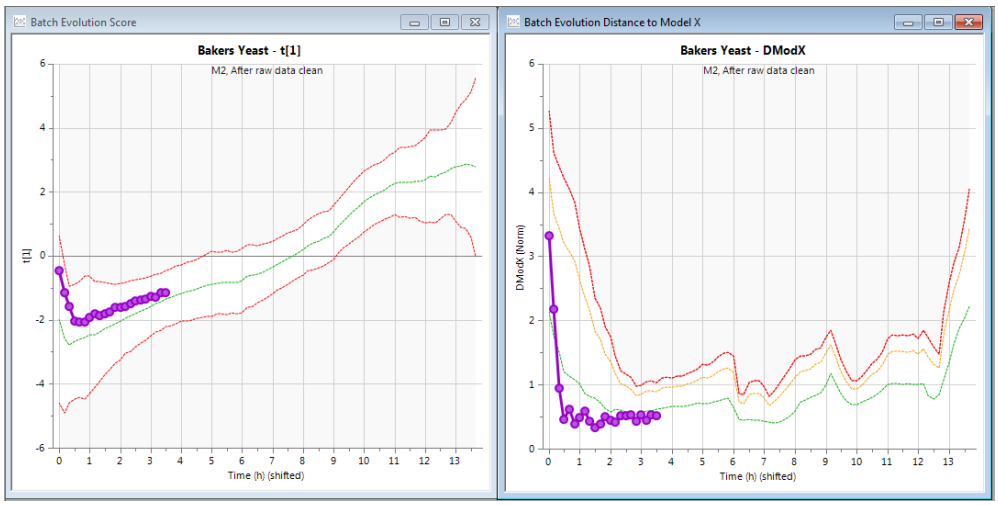


Figure 16.31: Two batch control charts of SIMCA-online showing the development of a good batch.

16.8.4 Related approaches

There are several methods related to the BSPC-approach presented here. Among these methods, we find multiway-PCA (MPCA) and multiway-PLS (MPLS) heavily promoted by MacGregor and coworkers [Nomikos and MacGregor, 1995a; Nomikos and MacGregor, 1995b]. MPCA and MPLS are extensions of PCA and PLS which can handle three-way batch data. These methods use unfolding of data preserving the batch direction.

An adaptive, recursive approach to updating of multi-block PCA and PLS models can be used [Rännar, *et al.*, 1998]. In this approach, each time slice of a three-way data table is treated as a block in a hierarchical model. This model is computed by a recursive updating, starting from the initial time slice. The result is a PCA model calculated such that for a batch under progress no filling in of future missing measurements is required. Limited tests of the hierarchical approach indicate its utility, and it usually needs fewer model dimensions than related approaches.

Another MPLS based approach, MPLS-DA, was used by Louwerse and colleagues [Louwerse, *et al.*, 1999] for modeling differences between parallel batch reactors in PVC polymerization. By incorporating contribution plots they were not only able to tell good and bad batches apart, but also to unveil unseen reactor differences.

Finally, we mention the work by Dahl and coworkers, aimed at comparing the performance of MPCA and PARAFAC (parallel factor analysis, see [Bro 1999]) in batch modeling [Dahl, *et al.*, 1999]. They used these two approaches to model batch data arising from manufacturing of a condensation product. They found both methods useful for modeling and monitoring batch process data. However, they found MPCA to be preferable in two respects, (i) its more widespread use and (ii) its ability to keep interpretation ambiguities to a minimum.

16.8.5 Batch duration and synchronization

One assumption of most methods used for the analysis of batch data is that batches have equal duration and are synchronized, i.e., measurements are made at the same time points for all batches. If this is not the case, the batch data need to be aligned. Wold and coworkers use stretched and contracted values of local time, and thus re-express each batch in terms of interpolated data so that in the end each batch has the same number of rows for the same time points. A subsequent analysis of the same type is then conducted, but using a dependent variable aligned according to predicted local time from the first analysis. Alternatively a maturity variable is used as the basis for batch synchronization instead of the local time. This approach is actually preferable since no stretching or contraction is needed, just interpolation to the same set of maturity levels for all batches. This maturity variable is a variable that expresses the degree of batch completion, e.g., alcohol content in beer

brewing, cell count or cell density in biotech fermentation, the amount of energy generated by an exothermic reaction, or the remaining amount of starting material in a reaction.

Another approach to batch synchronization based on dynamic time warping (DTW) was reported by the group of MacGregor [Kassidas, *et al,* 1998]. DTW has been used in speech recognition and is able to synchronize signals by appropriately translating, expanding, and contracting small trajectory segments. This method was found useful for batch alignment [Martin, *et al.*, 2006; Gonzalez-Martinez, *et al.*, 2011]. Correlation optimized warping (COW) is a related approach also proposed in this context [Tomasi, *et al.*, 2004; Fransson and Folestad, 2006].

17 Multivariate time series analysis

17.1 Objective

The objective of Chapter 17 is to describe multivariate time series analysis (MTSA) of dynamic systems. Time series analysis (TSA) is used to model the dynamics of a process, by relating the output variables (Y) to current and past values of the input variables (X) and/or past outputs.

17.2 Outline of Chapter 17

Chapter 17 is organized as follows:

1. Introduction to time series data
2. Review of time series analysis
3. Review of the transfer function
4. Illustration of multivariate time series analysis with 2 examples:

   a. The first example is a continuation of the analysis of the CUPRUM dataset (see Chapter 9). MTSA is here applied to the previously computed latent variables (principal component scores).

   b. The second example deals with manufacturing of pulp using a continuous Kamyr digester and using PLS time series analysis to model the Kappa number.

17.3 Introduction to time series data

17.3.1 Definition of time series data

A *time series* is a sequence of observations measured sequentially in time [Box, *et al.*, 1994]. Time series data are common in most fields of science, such as economics, business, natural sciences, and process monitoring and control. Any discrete process or regularly sampled continuous process can be represented by a time series. Sampling over time usually makes neighboring *observations* correlated. The nature of these correlations is important for understanding the process dynamics.

To quote the *Little Oxford Dictionary* (Clarendon Press, Oxford, 1994), the analysis of system dynamics involves a "mathematical study of motion and forces causing it". Hence, TSA is concerned with techniques for modeling and understanding dependency among observations, usually using stochastic models. A stochastic model calculates the probability that a future process value will lie between two specified limits by taking the "noise" (unknown factors disturbing the process) into account. In contrast, a deterministic model is used for the exact calculation of time dependent quantities, and assumes that all important factors are known.

Additionally, a time series is *stationary* when it is in equilibrium around a constant mean and with a constant variance.

17.3.2 Multivariate time series data

A limitation of classical time series analysis is that collinearities among measured *variables* are largely neglected. This is unfortunate because relationships among variables are as important as relationships among observations. As an illustration, consider the copper electrolysis (CUPRUM) example described in Section 9.4. We recall that eight impurities (variables) were measured for 730 copper samples, taken twice a day during one year of production. It was shown that these eight variables (impurities) clustered in two distinct groups, indicating a strong correlation structure. Univariate time series analysis of these data is therefore not appropriate. That would correspond to fitting one separate model for each individual impurity variable, and hence ignoring the interdependence of the variables. Multivariate time series analysis (MTSA) investigates and models both variables and observations, as well as their correlations, see e.g. [Lestander, *et al.*, 2012].

17.3.3 Objectives of multivariate time series analysis (MTSA)

The goal of multivariate time series analysis is the modeling of the system under dynamic conditions and the prediction of future values of the studied time series, based on its current and past readings.

17.3.4 Transfer functions and PLS time series analysis

When data are collected under transient or dynamic conditions (Figure 17.1), rather than at steady-state, the observations are no longer independent and regular PLS models relating the output Y to the input X are no longer adequate. To model a process under dynamic conditions, i.e., to model the transient behavior of the process, we use PLS time series analysis.



Figure 17.1: Process data are often collected under transient or dynamic conditions, rather than at steady-state. The quantities within brackets, x_t and y_t, can be vector variables, which organize the data as matrices. The objective of the data analysis is to build a dynamic model relating the transient behavior of Y to observed inputs X, past inputs, and/or past outputs.

17.4 Review of time series and their properties

17.4.1 Stationary stochastic models

The basic stochastic process model is the white noise sequence a_t (t = 1,2, etc), where a_t is a random variable with mean zero and constant variance. Time series models are based on the idea that a series z_t at time t = 1,2,3..., n, in which successive observations are correlated, can be regarded as generated by an uncorrelated series of shocks, or white noise, a_t, passing through a linear filter (Figure 17.2). This linear filter can be parsimoniously parameterized by the family of autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models of order (p,q) with a transfer function $\theta(B)/\phi(B)$, where

$\theta(B) = 1 - \theta_1(B) - \theta_2(B) ..... \theta_q(B)$ (the MA operator) (eqn. 17.1)

$\phi(B) = 1 - \phi_1(B) - \phi_1(B) ..... \phi_p(B)$ (the AR operator) (eqn. 17.2)

and B denotes the backshift operator $Bz_t = z_{t-1}$.

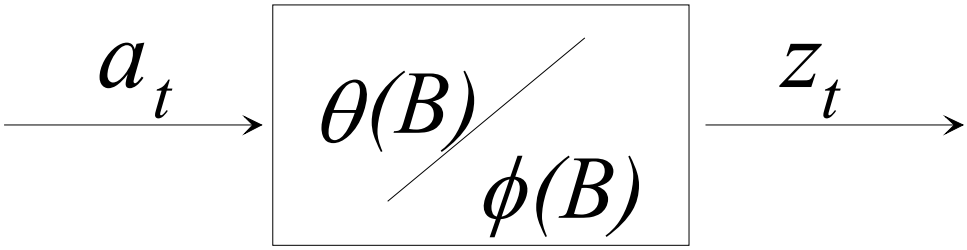


Figure 17.2: A linear filter transferring an uncorrelated series of white noise shocks, a_t, to a time series, z_t, in which successive observations are correlated. This linear filter can be parsimoniously parameterized by an ARMA model of order (p,q) with the transfer function θ (B) / ϕ (B).

The ARMA is stationary when the roots of the equation $\phi(B) = 0$ are greater than 1 in absolute value, i.e., lie outside the unit circle.

17.4.2 Autocorrelation and partial autocorrelation of the AR, MA, and ARMA models

The autocorrelation, r_L, of lag L is defined as:

$r_L = c_{zz}(L) / c_{zz}(0)$

$c_{zz}(L) = 1/N(\Sigma (z_t - z_{avg})(z_{t+L} - z_{avg}))$ Σ: $t=1$ to $N-L$

$c_{zz}(0) = 1/N(\Sigma (z_t - z_{avg})^2)$ variance z_t

for $L = 0,1, 2..$ etc

Note $r_L = r_{-L}$ (eqn. 17.3)

and is a measure of the linear dependence between adjacent observations. In these expressions, $c_{zz}(L)$ is the auto-covariance of z_t, $c_{zz}(0)$ the variance of z_t, N the number of observations, and L the lag.

- **Autocorrelation of an AR(p)**

$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ...+ = \phi_p\rho_{k-p}$ for $k > 0$ (eqn. 17.4)

The autocorrelation function of a stationary AR process of order *p***,** consists of a mixture of damped exponential and damped sine waves. The autocorrelation function of a stationary AR can be used to identify the order p of the AR model.

- **Partial autocorrelation function of an AR(p)**

$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ...+ \phi_p\rho_{k-p}$ for $k > 0$ (eqn. 17.5)

where ρ_{k-p} is the partial autocorrelation at lag p. The partial autocorrelation has a cut-off at lag p.

- **Autocorrelation of an MA**

The autocorrelation function of an MA(1):

$\rho_k = - \theta_1/ (1 + \theta_1^2)$ for $k =1$ (eqn. 17.6)

and $\rho_k = 0$ for $k = 2$. The autocorrelation function of an MA process of order 1 has a cut-off after lag 1.

$\rho_k = 0$ for $k = 2$. Autocorrelation function of an MA(q). The autocorrelation function of an MA model of order q has a cut-off after q lags.

- **Partial autocorrelation of an MA(q)**

The partial autocorrelation function of an MA behaves like the autocorrelation function of the corresponding AR-process.

17.4.3 Non-stationary stochastic models

Box and Jenkins have shown that some non-stationary time series (the mean or level of the series is *not* fixed) can be made stationary by taking differences (first or second) of

successive values of the series z_t. Hence, some non-stationary time series are modeled by the auto-regressive integrated moving average, ARIMA (p,d,q) model, where after taking the appropriate difference, d, an ARMA (p,q) model is fitted to the differenced series.

17.4.4 Identifying the time series model

In order to fit an ARIMA (p,d,q) model to the time series z_t we have to:

(1) Determine the degree of differencing, *d*, i.e., the number of successive differences needed to make the series z_t stationary. Usually d = 0 (the series is stationary) or 1 or 2.

(2) Identify the order (p,q) of the transfer function.

(3) Estimate the parameters of the ARMA model.

The principal tools used to estimate the parameters *d*, *p*, and *q* are the autocorrelation and partial autocorrelation functions.

The failure of the autocorrelation function to die out rapidly is an indication that a root of the AR operator is close to unity and suggests non-stationarity of the series z_t. The degree of differencing, *d*, is reached when the autocorrelation function dies out quickly. The appearance of the autocorrelation function of the appropriately differenced series provides clues about the orders of *p* and *q*, the auto-regressive and moving average operators [Box, *et al.*, 1994].

17.4.5 Modeling multivariate time series data

The recommended approach to modeling multivariate time series data is to:

(1) Analyze the raw data with PCA and extract a few scores.

(2) Identify the scores' time series properties using the autocorrelation and partial autocorrelation functions.

For the partial autocorrelation, one approach is to lag the PCA scores and fit PLS models where each score becomes a function of its lags. (Lagging is explained in Section 17.6.2. and Section 17.7.5). Here one should take as many PLS components as available (i.e., get the regression solution). The resulting PLS regression coefficients – corresponding to the maximum number of PLS components – are then proportional to the partial autocorrelation coefficients. Hence, these can be plotted as a function of the lag number to display the partial autocorrelations of the desired PCA score.

The autocorrelation function and the partial autocorrelation function are then used to identify the degree of differencing and the order of the series.

(3) Finally, this model is fitted.

For illustrative purposes, we will use lagging of latent variables and PLS modeling when identifying time series families in the CUPRUM application (see Section 17.6).

17.5 The transfer function and PLS time series analysis

17.5.1 Introduction

When data are collected under transient or dynamic conditions, rather than at steady-state, the observations are no longer independent and regular PLS models relating the output Y to the input X are no longer adequate. We then use PLS time series analysis to model the behavior of the process under dynamic conditions.

With a linear and stationary dynamic system (Figure 17.3), we can represent the relation between input and output variables when the system is at steady-state as:

$$Y_\infty = gX_\infty \qquad (g = \text{steady-state gain}) \qquad \text{(eqn. 17.7)}$$

When the input varies, Y_t and X_t measure the deviations from steady-state. The system inertia can be represented by a linear filter

$$Y_t = \mu + v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + ...+...+ n_t. \qquad \text{(eqn. 17.8)}$$

In other words, Y at time t is a function of X at time t and an infinite series of X at previous times t-1, t-2, t-3, etc.

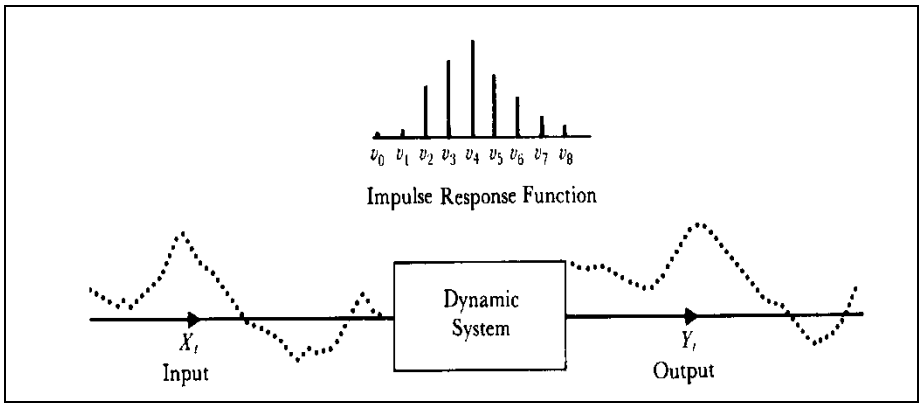


Figure 17.3: Schematic of a dynamic system. See text for further explanation.

17.5.2 The FIR Model

Using only a finite number of previous X´s leads to the *finite impulse response (FIR)* model. The requirement to be able to use an FIR model is that the whole dynamic response time is represented by past inputs.

If the maximum response time to any change in the input is γ, and T is the sampling time, then the number of regressors (lags) required in the model is γ / T. With responses having a slow settling time, this usually requires a large number of regressors.

17.5.3 Transfer function models

The output Y at time t can be represented by a transfer function model.

$$Y_t = [w(B)B^b / \delta(B)]\, X(t) + [\theta(B) / \varphi(B)]\, a(t) \qquad \text{(eqn. 17.9)}$$

This is the most general model with *a* complete models for the process and the noise, using past inputs and outputs (Figure 17.4).

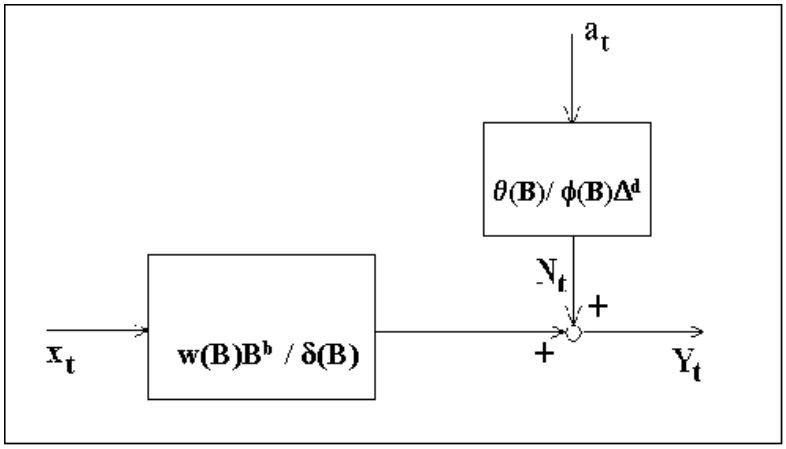


Figure 17.4: Illustration of transfer function concept.

This leads to the following model types:

- **ARX:** past inputs X(t-k) and past measured outputs Y(t-k). No separate model for the noise.
- **ARMAX:** past inputs X(t-k) and past measured outputs Y(t-k), and a partially separate model for the noise.

- **O(utput) E(rror):** past inputs X(t-k) and past predicted outputs Yhat(t-k), no noise model. Concentrates on the process dynamics.
- **Box and Jenkins:** complete models for the process and the noise, using past inputs and outputs.

17.5.4 Identification of transfer function model

To use a transfer function model, one needs to identify the order of the polynomials in the transfer function. A plot of the coefficients of the FIR model versus lag number is the tool to help in the identification of the order of the transfer function. We will give an example of modeling a Kamyr digester with an **ARX**, using first an **FIR** model to identify the order of the polynomials in the transfer function of the output.

17.6 MTSA and lagging of latent variables (Example: CUPRUM)

Wikström and coworkers have published two reports [Wikström, *et al.*, 1998a; Wikström, *et al.*, 1998b] regarding MSPC and MTSA of the CUPRUM dataset. In this section, we will review their findings regarding the MTSA.

17.6.1 PCA of original CUPRUM data

The PCA of the eight impurity variables yielded a two-component model explaining 67% ($R^2X = 0.67$) and predicting 40% ($Q^2X = 0.40$) of the sum of squares of X (see Section 9.4). The model interpretation was given in Section 9.4. The two PCA scores, t_1 and t_2, are linear combinations of the original variables and summarize their systematic information. Hence, we shall use these for monitoring purposes in conjunction with MTSA.

17.6.2 MTSA of the t_1-score

The autocorrelation of the score t_1 lagged 30 times is displayed below (Figure 17.5). This graph exhibits the exponential decay characteristic of an ARMA(1,1) model, when the root (coefficient) of the AR-operator is close to the unit circle (slow decay) [Box, et al., 1994].

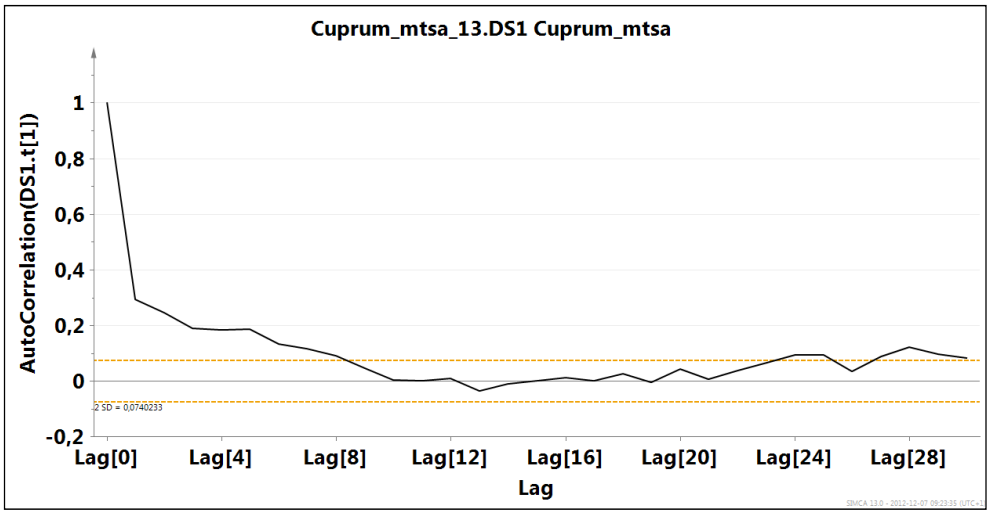


Figure 17.5: The autocorrelation function of the score t_1 plotted for 30 lags.

Next, the coefficients of the ARMA (1,1) model for the t_1-score were *estimated*. This was accomplished by calculating a PLS model in which the t_1-score was modeled as a function of two predictor variables, $t_{1,t-1}$ (its first lag) and a_{t-1} (the white noise at lag one). An estimate of the latter (a_t) was obtained by first fitting an AR(1) model to t_1 and using the resulting residual vector as an approximation for a_t. This is graphically illustrated in Figure 17.6.

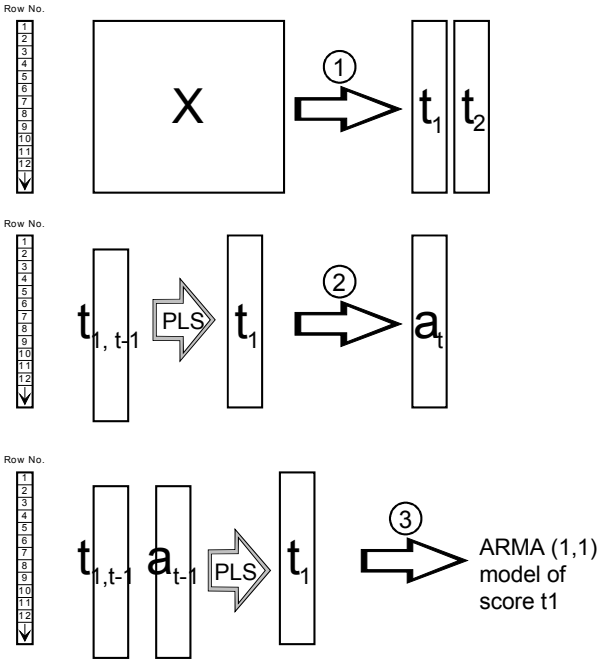


Figure 17.6: Illustration of the ARMA (1,1) modeling of the score t_1. (1) PCA of original data gives t_1 and t_2. (2) AR(1) modeling of t_1 gives an estimate of the white noise, a_t. (3) ARMA(1,1) modeling of t_1 using $t_{1,t-1}$ and a_{t-1} as predictors.

The PLS regression coefficients of the ARMA (1,1) model are displayed in Figure 17.7. As expected, the coefficient of $t_{1,t-1}$ is close to unity (coefficient = 0.83). Whenever the root of the AR operator approaches unity (as in this case), it is preferable to use a non-stationary model, in this case the ARIMA (0,1,1). (The first difference of the series is a moving average model of order 1).

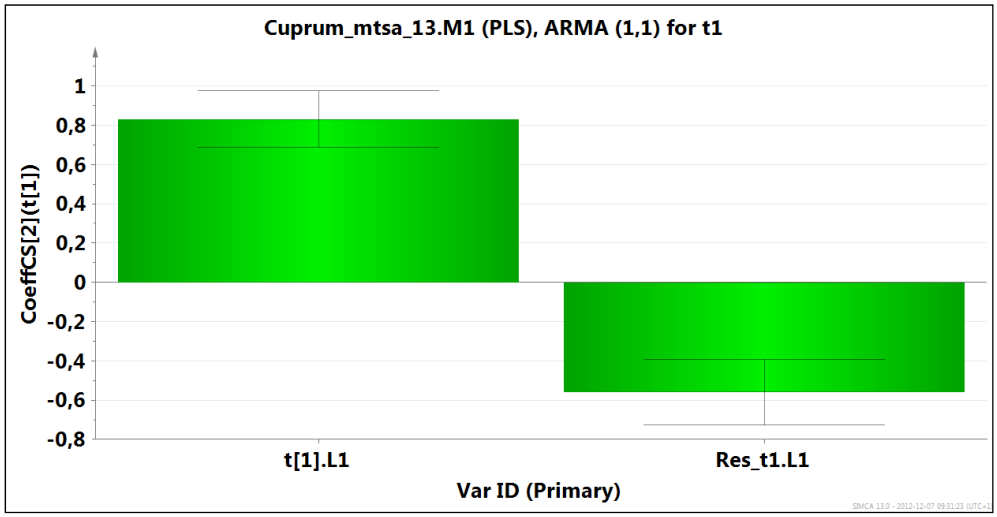


Figure 17.7: PLS regression coefficients of the ARMA (1,1) model fitted for t_1. The first coefficient relates to $t_{1,t-1}$ (i.e., t_1 lagged once) and the second to the white noise at lag 1 (a_{t-1}).

The ARIMA (0,1,1) model is an exponentially weighted moving average (EWMA) model [Box, *et al.*, 1994]. Thus, an EWMA control chart is recommended for monitoring t_1. Such a chart is useful for monitoring a dynamic process with memory and drift (non-stationary mean). The extent of memory is shown by the λ parameter ($\lambda = (1 + \theta)$; θ is the coefficient of the MA operator), ranging between 0 (infinite memory) and 1 (no memory). The value of

λ (in SIMCA), may be user specified, or estimated from the data as the value minimizing the predictive error sum of squares of one-ahead forecasts.

The value of the θ coefficient, as estimated from the ARMA (1,1) model, is ≈ -0.6 (Figure 17.7), pointing to a λ-value of around 0.4. However, when estimating the value of λ resulting in the lowest prediction error, the value 0.09 was obtained (see Section 15.6.3). We decided to set $\lambda = 0.2$ for all applications discussed in this chapter. This value is a reasonable compromise and corresponds to a rather long process memory.

Figures 17.8 and 17.9 show EWMA control charts of t_1 using subgroup size 1 and 2, respectively. The latter control chart is probably more appropriate since two copper samples were taken each day. One can see the smoothing due to subgroup size 2 (smoothing usually increases with subgroup size). It is evident that the process shows a drift over the whole year of production, and passes the upper and lower control limits (UCL, LCL) on several occasions. However, t_1 does *not* show a long-term periodic behavior.

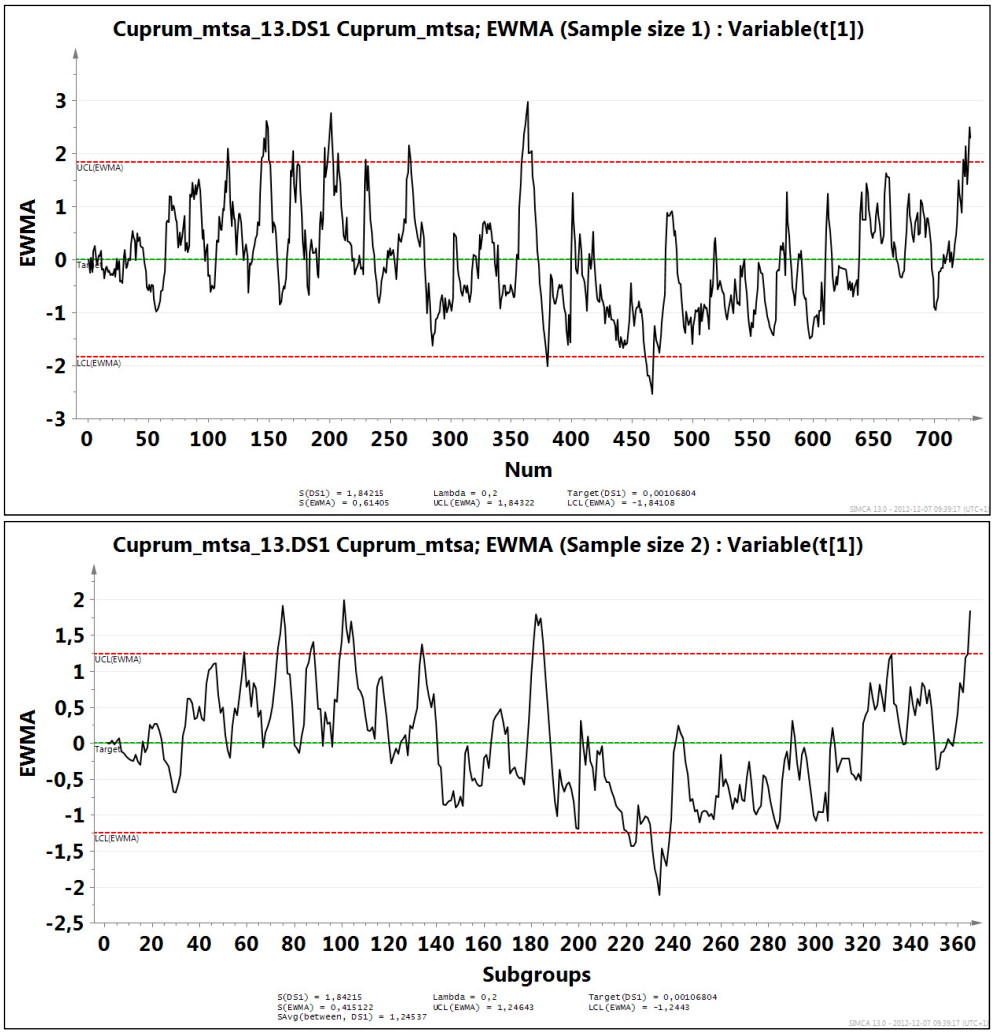


Figure 17.8: (above) A multivariate EWMA control chart for t_1 using no sub-grouping (i.e., sub-group size = 1) of the data and $\lambda = 0.2$. The upper and lower control limits (UCL, LCL) are displayed. The numerical order corresponds to sample number. Figure 17.9: (below) EWMA control chart of t_1 using sub-group size 2 and $\lambda = 0.2$. The numerical order corresponds to day number.

17.6.3 MTSA of the t_2-score

The dynamics of the t_2-score were investigated similarly to t_1. Thus, we first *identified* the time series family and again found ARMA (1,1) to be relevant. For reasons of brevity, no plots relating to this phase are shown. Secondly, we *fitted* an ARMA (1,1) model to t_2. The resulting PLS regression coefficients of the ARMA (1,1) model are shown in Figure 17.10.

Both regression coefficients are close to one, again indicating that the EWMA model is appropriate [Box, *et al.*, 1994].

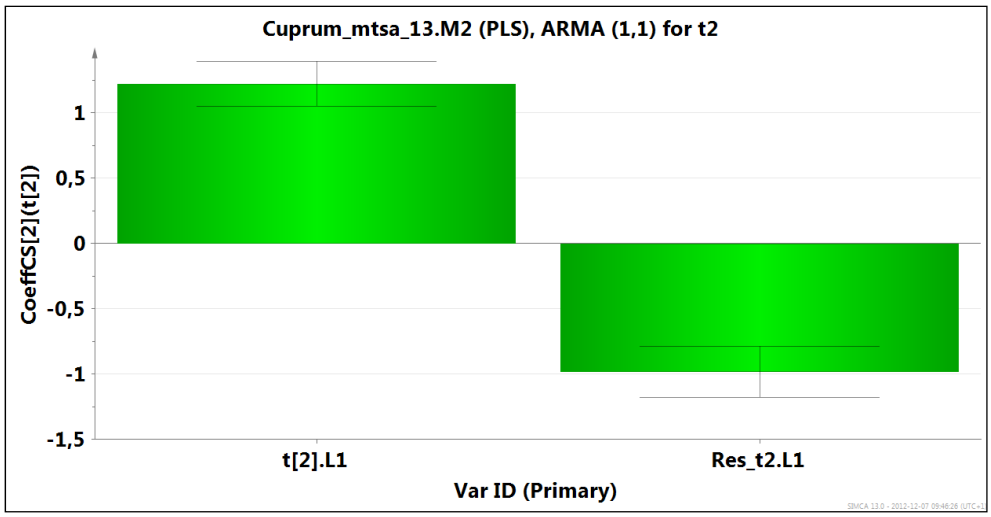


Figure 17.10: PLS regression coefficients of the ARMA (1,1) model fitted for t_2. The first coefficient relates to $t_{2,t-1}$ (i.e., t_2 lagged once) and the second to the white noise at lag 1 (a_{t-1}).

An EWMA control chart with subgroup size 2 and $\lambda = 0.2$ is plotted in Figure 17.11. This control chart shows that t_2 describes a "downward" process movement lasting for nearly 200 days, starting from the beginning of the sampling period. This sequence ends around day 225, whereupon a new and "upward" shift begins. The t_2-score does *not* describe a long-term periodicity.

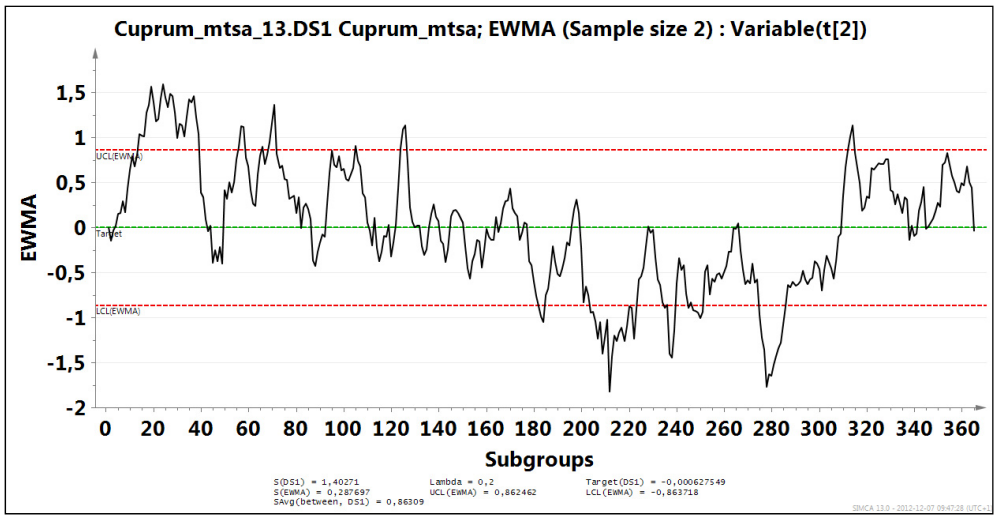


Figure 17.11: EWMA control chart of t_2 using sub-group size 2 and $\lambda = 0.2$. The horizontal axis represents the day number.

17.6.4 Summary of MTSA of CUPRUM data

Multivariate time series modeling was applied to an electrolysis process. Two PCA scores were examined and for both ARMA (1,1) models were found appropriate. When fitting these two ARMA (1,1) models, the roots (coefficients) of the AR-operator in both cases were close to unity (see Figures 17.7 and 17.10). Hence we decided to use an ARIMA (0,1,1) model, which, in turn, is an EWMA model [Box, *et al.*, 1994]. The EWMA control chart is therefore recommendable for monitoring t_1 and t_2.

The EWMA approach was discussed in Chapter 15 of this work and was compared with other control chart approaches. Here, it is sufficient to say that the EWMA approach often results in a smoothing of the data, making weaker process disturbances more obvious.

17.7 MTSA using PLS (Example: KAMYR)

17.7.1 The KAMYR digester process

The KAMYR dataset contains data from a pulp manufacturing process using a KAMYR digester [Dayal, *et al.*, 1994]. In this application, data were collected hourly during an eight month sampling campaign under closed loop. Figures 17.12 and 17.13 show time series plots of two measured variables. The fluctuating nature of the process is obvious.

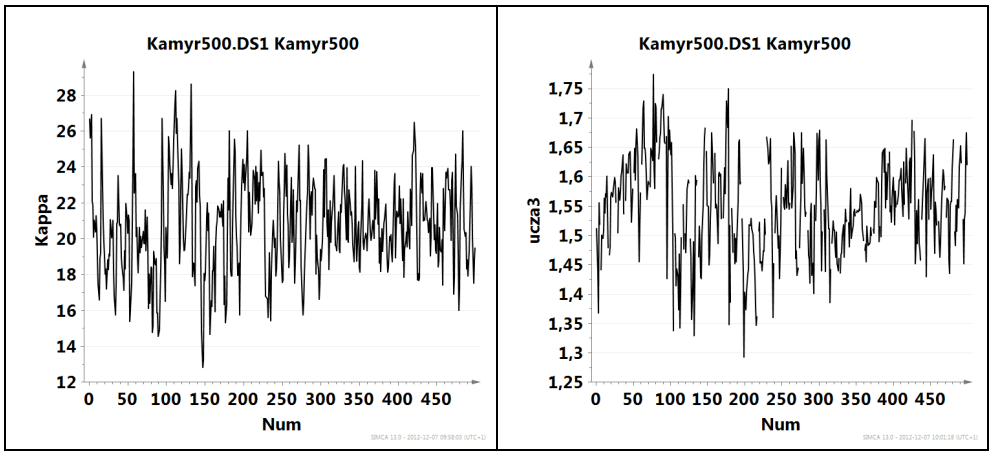


Figure 17.12: (left) Time series plot of Kappa number (Kappa) of pulp produced at 500 process time points.
Figure 17.13: (right) Time series plot of amount of active alkali in upper cooking zone (ucza) at 500 time points.

In this process, the incoming wood chips are fed into a steaming vessel. They fall into a chip chute submerged with white liquor, connected to a high-pressure feeder with staggered pockets. Chips and liquor are then introduced to the top of the digester and separated. The wood chips move through four zones, (i) impregnation, (ii) heating, (iii) cooking, and (iv) washing. During impregnation active alkali penetrate the chips. The heating set-up consists of two zones, employed to raise and control temperature. The cooking results in delignification, and washing is done to remove excess chemicals. At the bottom of the digester, pulp is continuously withdrawn and blow flow controlled by metering valves to maintain a constant chip level in the digester. The operating conditions of the digester are regulated according to the volumetric chip meter rate. The chip meter rate determines the retention time of the wood chips in the digester. All other material flows are adjusted to compensate for fluctuations in the chip feed rate.

During the production of pulp, delignification is an important step. Lignin can be regarded as the “glue” that binds cellulose fibers together in a wood matrix. The lignin must be removed to set the fibers free before paper production can take place. The target lignin content varies from application to application, e.g., for a white paper a low lignin pulp is desirable. The quality of the pulp, in terms of lignin content, is characterized by the Kappa number. A high Kappa number means high lignin content remaining in the pulp.

In the example used, 22 variables were measured, 21 X-variables and the Kappa number (see short explanation in Table 17.1). The data set used contains 500 observations (time points). The X-variables have been lagged 0, 2, 3, or 4 time shifts to account for delay (dead time). Thus, the representation ucza3 means that the variable active alkali in the upper cooking zone has been forward shifted 3 hours in relation to the Kappa number.

Table 17.1: Short description of X-variables.

Number	v-name	
1	Ykap	
2	chip4	chip meter rate
3	blcm	blow to cm ratio
4	blflw	blow flow
5	chpl4	digester chip level
6	uxt2	upper extr. temp
7	lxt2	lower extr temp 2
8	ucza3	upp. cook z. act alk 3
9	wlfl4	white liquor flow 4
10	wlst	white liq strength 4
11	aawd4	AA / wood ratio 4
12	chmo4	chip moisture 4
13	bsfl4	bin stream flow 4
14	lht3	low heater temp 3
15	uht3	upper heater temp 3
16	cmfr4	chip mass feed rate 4
17	tfflw	total filtr flow 0
18	xflw2	extraction flow 2
19	f18fw	f18 flow 0
20	stfl3	170* steam flow 3
21	tct4	top circul. Temp 4
22	lqs4	liquor sulphidity 4

17.7.2 The autocorrelation function

As with any type of data analysis, it is important in MTSA to plot the data to get a feeling for characteristic properties. Plotting of data is useful to find out how much dynamics there is in a system, if there is any periodicity or other cyclic behavior, and so on.

The autocorrelation function is one useful tool. It shows how a signal correlates with itself as a function of time shift (or "lag"). An autocorrelation function is symmetric around lag zero (no shifting in time), so the function is normally plotted one-sidedly for positive lags. Figure 17.14 shows a plot of the autocorrelation function of the Kappa number (Kappa) using 30 lags. The autocorrelation function seems to die out after lags 3 or 4. Similarly, Figure 17.15 shows the autocorrelation of ucza3. Lagging of variables is further described in Section 17.7.5.

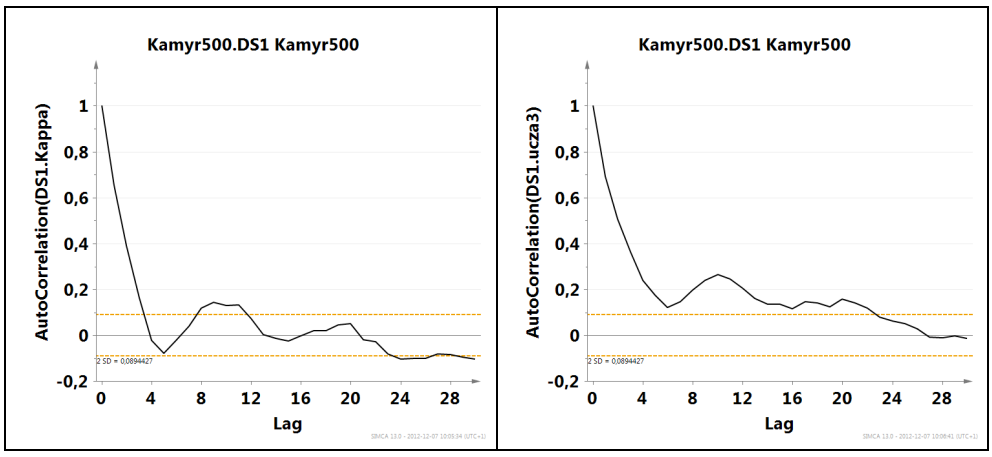


Figure 17.14: (left) Autocorrelation function of Kappa number up to lag 30. Figure 17.15: (right) Plot of autocorrelation of ucza3 using 30 lags.

The autocorrelation function is indicative of the nature of a time series, i.e., whether it is *stationary* or not. A time series is said to be stationary when the process is in equilibrium around a constant level or mean. Stationarity thus implies that the process/system average, or current level, is independent of the time *t*. A stationary process is characterized by an autocorrelation function that dies out quickly. The autocorrelation function of a non-stationary process does not die out quickly. In this example the autocorrelation function of the Kappa number (Figure 17.14) dies out quickly, thus suggesting that the process is stationary in this quality variable.

17.7.3 The cross-correlation function

The cross-correlation function is another exploratory tool. It assesses how one reference time series correlates with another time series, or several other series, as a function of time shift ("lag"). Also, note that in contrast to an autocorrelation function, a cross-correlation function is typically not symmetrical around the zero´th lag.

Figure 17.16 displays the cross-correlation function between the reference variable Kappa number and the ucza3 X-variable. Here the maximum correlation is -0.6 located at 0 hour lag. This is because ucza3 has already been shifted 3 hours to take account of the delay or dead time. Similarly, Figure 17.17 shows another cross-correlation involving the Kappa number as the reference variable. Here, however, the X-variable (chmo4) does not correlate with the reference variable. (Note that Figures 17.16 and 17.17 are scaled in the same way.)

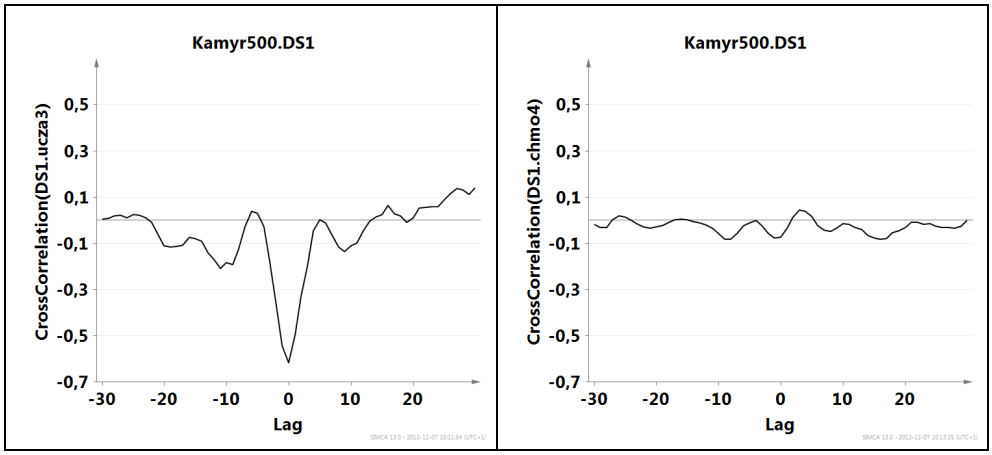


Figure 17.16: (left) Cross-correlation plot between the Kappa number and ucza3. At lag 0 there is a correlation peak, and the correlation coefficient is –0.6. Figure 17.17: (right) Cross-correlation function between the Kappa number and chmo4.

17.7.4 PLS modeling of 21 X-variables and the Kappa number

The next step in the data analysis is PLS. According to cross-validation, a two-component model resulted, which utilized 34% of X ($R^2X = 0.34$) to model 48% ($R^2Y = 0.48$) and predict 45% ($Q^2Y = 0.45$) of the Kappa number.

Figure 17.18 shows the inner relation in the first component. The observations are somewhat scattered around the ideal diagonal running from the lower left-hand to the upper right-hand corner. However, there is a sound correlation and we have no outliers in the relation between X and Y (expressed by t and u).

Figure 17.19 provides a scatter plot of the PLS weights. These tell us that three variables, ucza3 (X-variable no 7), tct4 (X-variable no 20), and chpl4 (X-variable no 4) are more influential for the Kappa number than other variables.

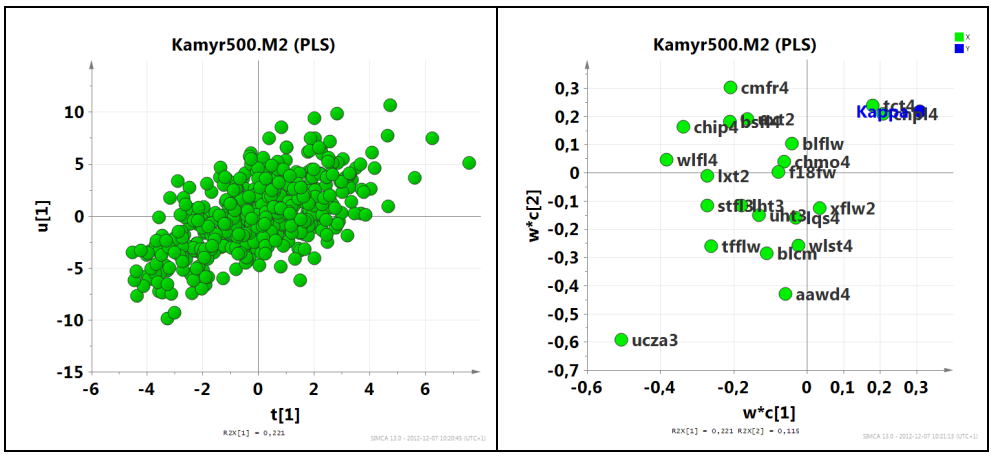


Figure 17.18: (left) PLS t_1/u_1 score plot of the overview KAMYR model. This plot informs us that there is a correlation between X and the Kappa number. Figure 17.19: (right) PLS weights showing how the X-variables influence the Kappa number. The X-variables ucza3, tct4 and chpl4 are most important.

The above PLS model was developed assuming a steady-state situation and does not take account of the dynamics of the process. We will now develop a PLS time series model of the Kamyr digester to take into account its transient behavior.

17.7.5 PLS time series analysis: Lagging of variables

In order to accomplish PLS-TSA *lagging* of data is necessary. Lagging implies that new variables are created from the old ones by forward shifting data (Figure 17.20). Variables belonging to both the X- and the Y-matrices can be expanded with lagged variables. Thus, by means of lagging we may generate an X-matrix consisting of [Y_{t-1}, …,Y_{t-L}, X_t, X_{t-1}, …, X_{t-L}] where L is number of lags. This allows us to model how the current process situation at time t is influenced by the process variables at L time units earlier. A variable can be lagged with several different lags, e.g., 2, 4, 12, 24 and 36. Also, depending on how we define the PLS X-block we can generate different types of time series models (see discussion in Section 17.5.3).

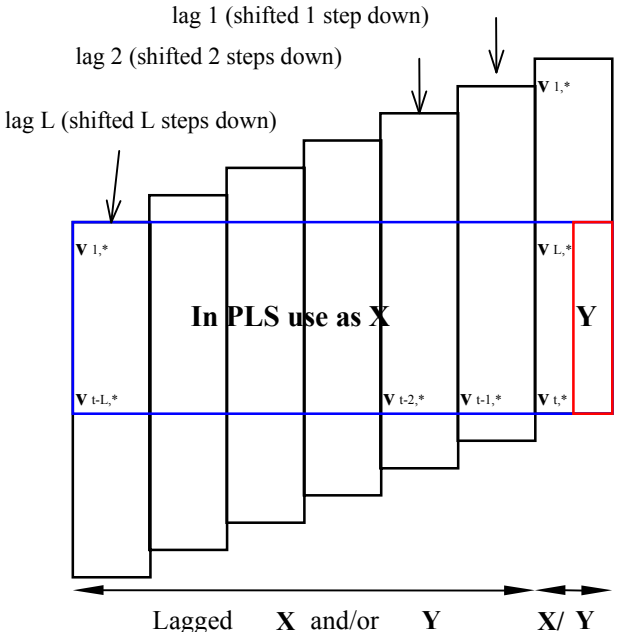


Figure 17.20: Overview of the lagging phase. Both X- and Y-data can be lagged and analyzed with PLS. Expanded X- and Y-variables form the new X-matrix. Original Y-variables are still used as Y-data. The resulting data table becomes irregular, and is regularized by dropping rows from the top and/or bottom; i.e. by omitting a certain amount of data.

17.7.6 PLS finite impulse response (FIR) model of Kappa number

The FIR-model included 20 lags of all the X's. This was deemed sufficient in view of the slow sampling time, i.e. 1 hour. Thus, we added 21*20 = 420 new variables to the X-block. Because the maximum lag was 20, the first 20 rows of the data table were excluded, thus leaving 480 observations.

The new PLS-FIR model was better than its predecessor (Figure 17.21), using 36% of X ($R^2X = 0.36$) to model 58% ($R^2Y = 0.58$) and predict 47% ($Q^2Y = 0.47$) of the variation in the Kappa number. This improved model quality is also evident in the PLS t_1/u_1 score plot in Figure 17.22.

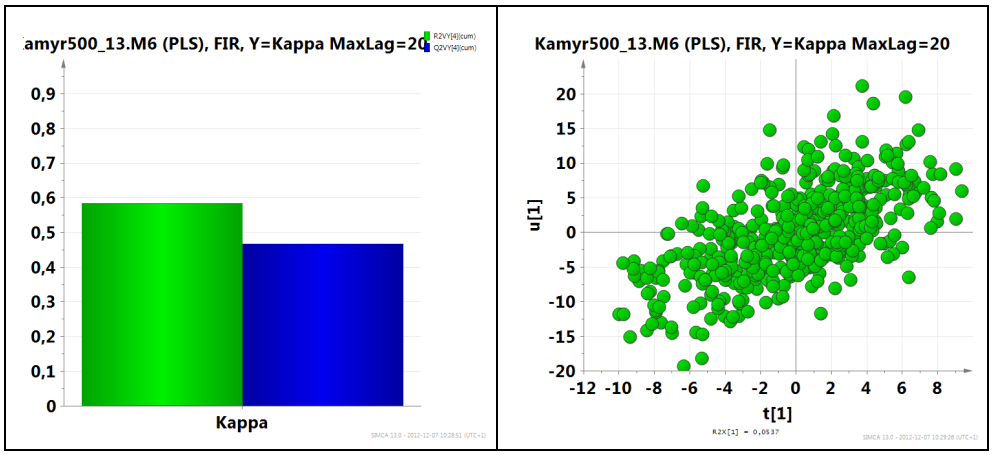


Figure 17.21: (left) Summary of fit plot of the PLS-FIR model based on lagging X according to the sequence 1-20.
Figure 17.22: (right) PLS t_1/u_1 score plot, indicating that a stabilization took place when lagging data.

Plotting the coefficients of the most important variable ucza3 as a function of lag helps in identifying the order of the transfer function model (see Step Response Plot in Figure 17.23).

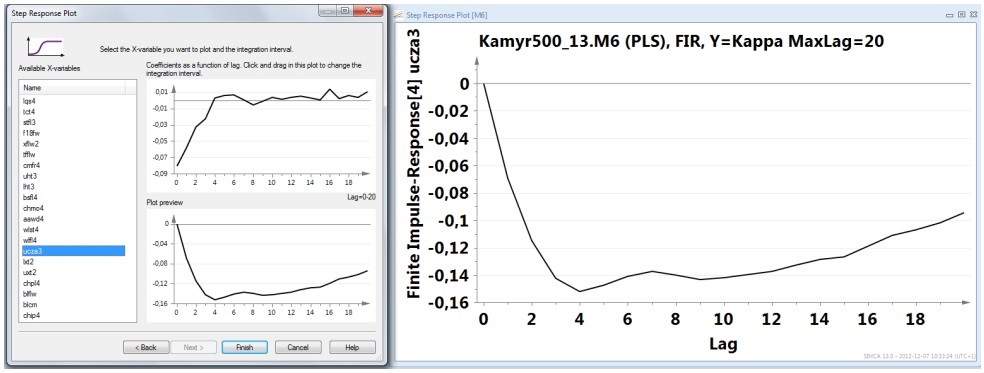


Figure 17.23: Step Response Plot addressing behavior of ucza3 as a function of lag number. This plot indicates that there is no need for further lagging of ucza3.

This plot (Figure 17.23) is typical of a first order exponential decay, and suggests that, to model the Kappa number, we should use a transfer function which is first order in Y and 0-order in X.

17.7.7 PLS transfer function model of Kappa number

To fit a first order in Y transfer function, we need now only lag the Kappa number by one lag. The model thus consists of all the X's and Y (Kappa number) lagged once.

The PLS modeling yielded a three-component transfer function model with $R^2X = 0.43$, $R^2Y = 0.60$, and $Q^2Y = 0.56$ (Figure 17.24). The PLS t_1/u_1 score plot (Figure 17.25) shows a good correlation between the X-variables and the Kappa number. There are no strong outliers in the relation between X and Y.

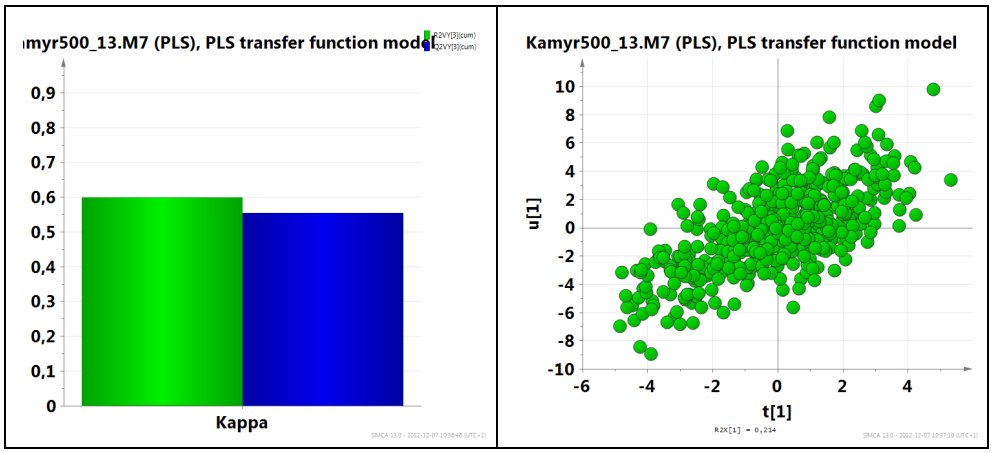


Figure 17.24: (left) Summary of fit plot of the refined PLS time series model. The model statistics are $R^2X = 0.43$, $R^2Y = 0.60$, and $Q^2Y = 0.56$. Figure 17.25: (right) PLS t_1/u_1 score plot of the refined PLS model. There are no strong outliers in the relation between X and Y.

In order to interpret the PLS transfer function model we consider the PLS weights displayed in Figure 17.26. From this plot we conclude that there are four important phenomena influencing the Kappa number:

1. The dynamics modeled here by the Kappa number lagged one hour.
2. The alkali in the upper cooking zone (adjusted 3 hours for delay), ucza3.
3. The group of variables on the upper left of the plot representing throughput variables.
4. tct4 (the upper level temperature) and chpl4 (the chip level in the digester).

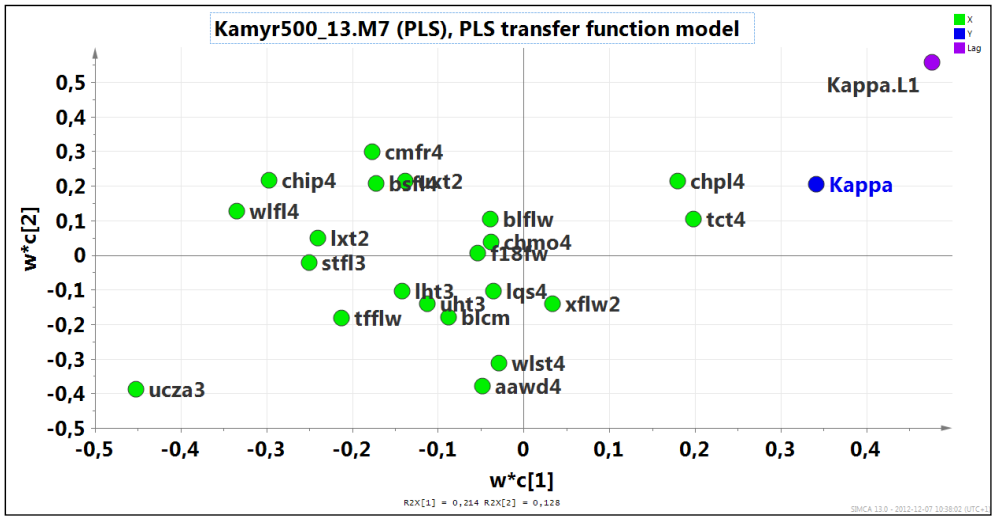


Figure 17.26: PLS weight plot of the transfer function model.

17.7.8 Summary of MTSA of the KAMYR data

PLS time series analysis represents a straightforward extension of conventional time series analysis. By lagging X- and Y-variables it is possible, in a multivariate way, to model the influence of past process readings on the current situation at time *t*. By first making an overview model to identify the appropriate lag, or lag sequence, of each individual variable, and then dropping the variables with the smallest regression coefficients, it is possible to create a time series model extracting relevant information about the past and its influence on the current process condition.

In the KAMYR dataset, there is no doubt that lagging of data is crucial for obtaining a predictive model of the Kappa number. The first lag of the Kappa number and ucza3 dominate the model, but the influences of throughput variables and tct4 and chpl4 are also important.

17.8 Questions for Chapter 17

1. What is a dynamic system?
2. What is a time series?
3. What are the characteristics of a stationary process?
4. How does one identify the ARIMA models ?
5. What are the objectives of multivariate time series analysis?
6. What is the autocorrelation function? How can it be used?
7. What is the cross-correlation function? How can it be used?
8. What is lagging of variables?
9. What is a FIR-model?
10. How does one proceed to identify and fit a transfer function model?

17.9 Summary and discussion

Univariate time series analysis involves the modeling of correlations among *observations*, while ignoring correlations among *variables*. Multivariate time series analysis seeks to overcome this by simultaneously considering correlations both between observations and between variables. This gives a more realistic picture of the studied process.

We have here tried to demonstrate how the standard multivariate techniques PCA and PLS, together with lagging of measured or latent variables, neatly create a multivariate time series analysis framework. This framework includes certain key steps.

- The first step comprises the summary of the originally measured variables into a few principal component scores. Because such scores are limited in number and mathematically orthogonal they are well suited for process monitoring purposes using control chart techniques. They are also useful for time series modeling.
- The second step in the framework consists of identifying which type of time series model is adequate for the PCA scores.
- The third step is the actual estimation of the appropriate time series model for each score. Several types of time series models can be fitted to describe the dynamics of a process. Some of these were indicated in Section 17.5.3.

The interesting aspect is that PLS brings the simpler variants of time series analysis back to the standard multivariate data analytical framework. The addition of lagged Y-data to the X-matrix often makes it possible to approximate Y with fewer regressors, because much of current Y-behavior is contained in past measurements of Y. An introduction to system identification methods for multivariate calibration and process modeling has been presented by Ergon [Ergon, 1998].

18 Process Analytical Technology (PAT) and Quality by Design (QBD)

18.1 Objective

This chapter introduces the concepts of Process Analytical Technology (PAT) and Quality by Design (QBD) and discusses their merits and possibilities within the pharmaceutical industry. PAT is founded on chemometric cornerstones like *m*ultivariate *d*ata *a*nalysis (MVDA), *m*ultivariate *s*tatistical *p*rocess *c*ontrol (MSPC) and *b*atch *s*tatistical *p*rocess *c*ontrol (BSPC). A major benefit of PAT is that of using multivariate process data to provide more information and a better understanding of manufacturing processes than conventional approaches (such as univariate SPC). Another benefit is the reduction of idle time at the end of a process step when product is often held awaiting laboratory results. This second benefit is a concept called "parametric release". QBD is founded on similar tools to PAT, but in addition emphasizes the need for *d*esign *o*f *e*xperiments (DOE) to achieve trustworthy results. A major benefit of QBD is the development of a sound scientific basis for a process design space that accommodates a range of defined variability in commercial process materials and operations and still produces the right product quality outcomes.

18.2 Introduction

In Chapters 15 – 17 we have seen how tools for multivariate data analysis (MVDA) and multivariate statistical process control (MSPC) form an integrated framework for process modeling, monitoring and optimization. Using this framework, processes that are running in a sub-optimal way can be better characterized and understood, and eventually be brought into a better state of control and maintenance. In this chapter, the objective is to deepen the discussion on process modeling, monitoring and optimization with special attention to the influence arising from regulatory activity and guidance. To this end, we shall consider two concepts denoted PAT and QBD. PAT is short for process analytical technology and QBD is short for quality by design. The utility of QBD and PAT is well documented in literature [Andersson, *et al.*, 2005; Luukkonen, *et al.*, 2008; Altan, *et al.*, 2009; Streefland, *et al.*, 2009; Laursen, *et al.*, 2010; Peterson, 2010; Wu, *et al.*, 2011; Xu, *et al.*, 2012; Kawabe, *et al.*, 2013; Macedo, *et al.*, 2013; Rozet, *et al.*, 2013].

18.2.1 The PAT initiative

PAT has its origins in an initiative of the US Food and Drug Administration, FDA, first launched in mid-2002. The goal of PAT is to improve the understanding and control of the entire pharmaceutical and biopharmaceutical manufacturing process. One way of achieving this is through timely measurements of critical quality and performance attributes of raw and in-process materials and processes, combined with multivariate data analysis (MVDA). This needs to be coupled with Design of Experiments (DOE) to maximize the information content in the measured data. A central concept within the PAT paradigm is that quality should arise as a result of design-based understanding of the processes, rather than merely by aiming to generate products that meet minimum criteria within defined confidence limits, and rejecting those that fail to meet the criteria.

There are many objectives associated with the PAT concept, but the foremost goal is to improve the understanding and control over the entire pharmaceutical or biopharmaceutical

manufacturing process. Briefly, PAT can be understood as a framework of tools and technologies for accomplishing this goal. Interestingly, the US FDA defines process understanding as: “the identification of critical sources of variability, management of this variability by the manufacturing process, and the ability to accurately and reliably predict quality attributes”. MVDA tools are clearly needed to achieve this, together with insight, process knowledge, and relevant measurements.

A common question is “what can be accomplished with PAT?” The answer to this is multi-faceted. In part, this is because PAT means different things to different people. Its answer may also depend on whether one is involved in academia, industry or a regulatory agency. Moreover, the connotations associated with PAT depend to a great extent on the context, e.g. whether it is to be applied to an existing process (continuous or batch) or a process still under development.

MVDA and PAT may be applied to important unit operations in the pharmaceutical industry. In our experience, the most common applications of PAT relate to on-line monitoring of blending, drying and granulation steps. Here, the measurement and analysis of multivariate spectroscopic data are of central importance. These spectroscopic data form the X-matrix, and if there are response data (Y-data), the former can be related to the latter using PLS or OPLS to establish a multivariate calibration model (a so called soft sensor model). However, MVDA should not be regarded as the only important approach; DOE should also be considered. DOE is very useful for defining optimal and robust conditions in such applications.

18.2.2 What are the benefits of using DOE?

DOE is complementary to MVDA. In DOE, the findings of a multivariate PCA, PLS, OPLS or O2PLS model are often used as the point of departure because such models highlight which process variables have been important in the past. Systematic and simultaneous changes in such variables may then be induced using an informative DOE protocol. In addition, the controlled changes to these important process factors can be supplemented with measurements of more process variables which cannot be controlled. Hence, when DOE and MVDA are used in conjunction, there is the possibility of analyzing both designed and non-designed process factors at the same time, along with their putative interactions. This allows us to see how they jointly influence key production attributes, such as the cost-effectiveness of the production process, and the amounts and quality of the products obtained.

Generally, DOE is used for three main experimental objectives, regardless of scale. The first of these is screening. Screening is used to identify the most influential factors, and to determine the ranges across which they should be investigated. This is a straightforward aim, so screening designs require relatively few experiments in relation to the number of factors. Sometimes more than one screening design is needed.

The second experimental objective is optimization. Here, the interest lies in defining which approved combination of the important factors will result in optimal operating conditions. Since optimization is more complex than screening, optimization designs demand more experiments per factor. With such an enriched body of data it is possible to derive quadratic regression models.

The third experimental objective is robustness testing. Here, the aim is to determine how sensitive a product or production procedure is to small changes in the factor settings. Such small changes usually correspond to fluctuations in the factors occurring during a “bad day” in production, or the customer not following the instructions for using the product.

The great advantage of using DOE is that it provides an organized approach, with which it is possible to address both simple and tricky experimental and production problems. The experimenter is encouraged to select an appropriate experimental objective, and is then guided through devising and performing an appropriate set of experiments for the selected objective. It does not take long to set up an experimental protocol, even for someone unfamiliar with DOE. What is perhaps counter-intuitive is that the user has to complete a whole set of experiments before any conclusions can be drawn.

The rewards of DOE are often immediate and substantial, for example higher product quality may be achieved at lower cost, and with a more environmentally-friendly process performance. However, there are also more long-term gains which might not be apparent at first glance. These include a better understanding of the investigated system or process, greater stability, and greater preparedness for facing new, hitherto unforeseen challenges.

18.2.3 QBD and Design Space

The importance of DOE to the pharmaceutical and biopharmaceutical industry is underlined by the growing interest in the Quality by Design (QBD) concept inspired by ICH. ICH, the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, is a collaborative body bringing together the regulatory authorities and pharmaceutical industry of Europe, Japan and the US to discuss scientific and technical aspects of drug registration. ICH's mission is to achieve greater harmonization to ensure that safe, effective, and high quality medicines are developed and registered in the most resource-efficient manner. ICH regularly releases guidance documents relating to quality, safety and efficacy. A number of guidelines on the Quality side, denoted ICH Q8, Q9 and Q11 (Q for Quality), concern methods and documentation principles of how to describe a complex production process or analytical system.

The ICH guidance documents position QBD based on DOE as a systematic strategy of addressing pharmaceutical development. It should be appreciated that it is both a drug product design strategy and a regulatory strategy for continuous improvement. QBD is based on developing a product that meets the needs of the patient and fulfils the stated performance requirements. QBD emphasizes the importance of understanding the influence of starting materials on product quality. Moreover, it also stresses the need to identify all critical sources of process variability and to ascertain that they are properly accounted for in the long run.

As mentioned above, the major advantage of DOE is that it enables all potential factors to be explored in a systematic fashion. Let us take, as an example, a formulator, who will be operating in a certain environment, with access to certain sets of equipment and raw materials that can be varied across certain ranges. With DOE, the formulator can investigate the effect of all factors that can be varied and their interactions. This means that the optimal formulation can be identified, consisting of the best mix of all the available excipients in just the right proportions. Furthermore, the manufacturing process itself can be enhanced and optimized in the same manner. When the formulation recipe and the manufacturing process have been screened, optimized and fine-tuned using a systematic approach based on DOE and MVDA, issues such as scale-up and process validation can be addressed very efficiently because of the comprehensive understanding that will have been acquired of the entire formulation development environment.

Thus, one objective of QBD is to encourage the use of DOE. The view of this approach is that, once appropriately implemented, DOE will aid in defining the design space of the process. The design space of a process is the largest possible volume within which one can vary important process factors without risking violation of the specifications (the demands on the responses). Outside this window of operation, there can be problems with one or more attributes of the product.

18.2.4 MVDA/DOE is needed to accomplish PAT/QBD in Pharma

Because the pharmaceutical and biopharmaceutical industry has complex manufacturing processes, MVDA and DOE have great potential and applicability. Indeed, the pharmaceutical sector has long been at the forefront of applying such technology. However, this has mainly been at the laboratory and pilot plant scale and has historically been much less widespread in manufacturing. The main explanation for this dichotomy is that the pharmaceutical industry has been obliged to operate in a highly regulated environment. This environment has reduced the opportunity for change which in turn has limited the applications of MVDA and DOE in manufacturing.

A change to this approach involves confidence in making changes, and those changes must be based on process and product understanding. The objective in modeling and analysis of these data is to develop process understanding. The FDA guidance on process analytical technology (PAT) provides four types of tools for generation and application of process understanding, including

- multivariate tools for design, data acquisition and analysis;
- process analyzers;
- process control tools, and
- continuous improvement and knowledge management tools.

In the FDA's description, multivariate methods represent a class of analysis methods, process analyzers are metrics that describe the state of a system and process control tools include techniques for monitoring and actively manipulating the process to maintain a desired state. The inclusion of continuous improvement and knowledge management tools stresses the importance of integrated data collection and analysis procedures throughout the life-cycle of a product. If process understanding is demonstrated through the use of these four PAT tools, the FDA offers less restrictive regulatory approaches to manage change, thus providing a pathway for integration of innovative manufacturing methods.

Pharmaceutical processes are complex with potential product variability due to both variation in operating conditions and raw materials. Final drug quality is influenced by everything that happens during manufacturing – each process step, each ingredient, the condition of the equipment and even subtle changes in the manufacturing environment itself can lead to variations in product quality. The univariate specifications presently used for characterization of raw materials cannot adequately describe their quality or influence on the final product, often allowing problems to go undetected. Only by developing multivariate models that account for each potential cause of variability, and by applying process analytics, can drug manufacturers establish a foundation for manufacturing quality systems at their facilities. MVDA, MSPC, BPSC and DOE should therefore be important elements of any PAT program, MVDA being the indispensable cornerstone.

With the strict quality demands on pharmaceutical products, product variability must be kept acceptably small, and consequently there is a need for *process control techniques* to keep critical process variables on specified trajectories, and *process monitoring* ensuring that all parts of the process remain inside specified "trajectory volumes" formed by the data in multidimensional space. In this context, process monitoring is sometimes referred to as process status visualization (PSV) and the trajectory volumes are regarded as desirable batch trajectories. Such desirable batch trajectories can easily be specified using multivariate parameters, such as, scores, Hotelling´s T^2, DModX, contributions, etc., which was discussed in Chapters 15 and 16.

18.2.5 Organization of Chapter 18

The remaining part of Chapter 18 is organized as follows. In the next section we present a DOE case study of a continuous process, a mineral sorting plant. The objective of showing this case study is to discuss the basic principles of DOE and to demonstrate how a design space can be searched for within the QBD paradigm. With this DOE discussion in mind, the next few sections are devoted to a discussion of PAT and two PAT case studies are given. The chapter ends with discussions and conclusions, including an introduction to how multivariate process models and other design space models can be used in process control.

18.3 A process optimization case study: SOVRING

In this section, we will study a process optimization investigation called SOVRING. Optimization is used after screening. The objective is (i) to predict the response values for all possible combinations of factors within the experimental region, and (ii) to identify an optimal experimental point. However, when several responses are treated at the same time, it is usually difficult to identify a single experimental point at which the goals for all

responses are fulfilled, and therefore the final result often reflects a compromise between partially conflicting goals.

18.3.1 The SOVRING dataset

Recall that the data analysis of the SOVRING dataset was presented in Section 9.6. We remember that three controllable factors were varied according to a central composite design in 17 experiments. These factors were feed rate of iron ore (Ton_In), speed of first magnetic separator (HS_1), and speed of second magnetic separator (HS_2). Among the 17 experimental runs, eight experiments arose from the factorial part of the design, six from locating experiments on the three factor axes, and three from doing replicated trials at the design center.

One way to study the shape of the resulting experimental design consists of plotting the raw data. Figure 18.1 shows a scatter plot of two of the designed factors, HS_1 and HS_2. The regularity of the points is apparent. Hence, the influence of the underlying design cannot be mistaken. Similar scatter plots can be generated for HS_1 versus Ton_In, and HS_2 versus Ton_In, but such pictures are not given here. It is interesting that the multivariate data analysis – PLS modeling carried out in next section – recognizes and utilizes this regularity expressed by the X-matrix (Figure 18.2).

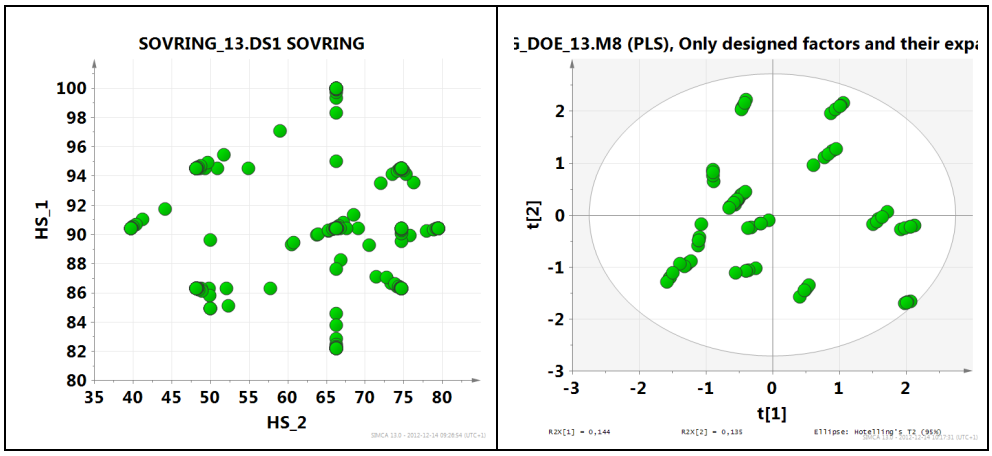


Figure 18.1: (left) Scatter plot of HS_1 against HS_2. Figure 18.2: (right) PLS t_1/t_2 score plot. Note the resemblance to Figure 18.1.

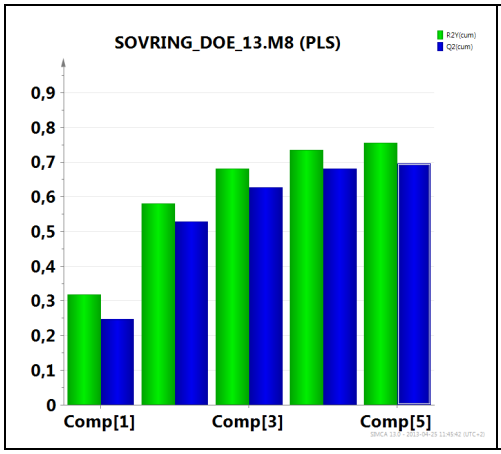
Each factor combination in the SOVRING design was preserved for 20 minutes to better capture normal process variation. The impact on manufacturing performance of each experimental point in the central composite design was encoded in the registration of six response variables (= 6 Y-variables). The concentrated material is divided in two product streams, one called PAR and another called FAR. The amounts of these two products form two important responses (they should be maximized). The other important responses are the percentages of P in FAR (%P_FAR), and of Fe in FAR (%Fe_FAR), which should be minimized and maximized, respectively.

Five observations with complete Y-data were registered for each factor setting. Thus, the SOVRING sub-set with complete Y-data contained 85 observations. This is the dataset considered in the next section.

18.3.2 PLS modeling of 85-samples SOVRING subset

A PLS model was established in the three designed factors (Ton_In, HS_1, HS_2) plus their six expanded terms. This model contained five components and utilized 61% of X (R^2X = 0.61) for modeling 76% (R^2Y = 0.76) and predicting 70% (Q^2Y = 0.70) of the response variation. Figure 18.3 displays the evolution of R^2Y and Q^2Y as a function of increasing model complexity. Figure 18.4 displays how well each individual response variable is modeled and predicted after five components. We can see that the two important responses

PAR and FAR are excellently described and predicted by the model. These two responses should be maximized. The other two important responses, %P_FAR and %Fe_FAR, which should be minimized and maximized, respectively, are fairly well modeled.



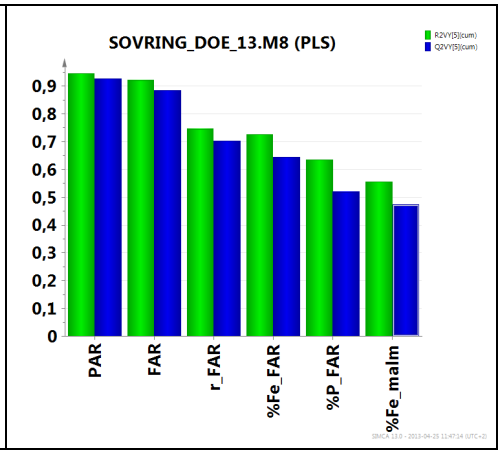
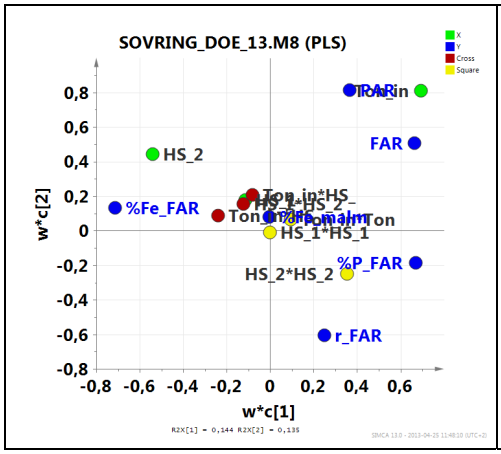


Figure 18.3: (left) Summary of fit plot of the PLS model of the SOVRING sub-set with complete Y-data. Five components were significant according to cross-validation. Figure 18.4: (right) Individual R^2Y- and Q^2Y-values of the six responses. The most important responses are PAR, FAR, %Fe_FAR, and %P_FAR.

Here, it would be pertinent to plot the relationships between pairs of t_a/u_a scores. However, these are very similar to Figure 9.24 in Section 9.6.4. We already know that the correlation between X and Y is sound in this application, and hence it is superfluous to provide more PLS score plots.

Figures 18.5 and 18.6 give scatter plots of the PLS weights. The two *amount of product* responses (FAR and PAR) are much influenced by the load (Ton_In), which is logical. However, what is really interesting is that the two *quality of product* responses (%Fe_FAR and %P_FAR) are strongly influenced by the operating performance of the second magnetic separator (HS_2). Hence, by adjusting HS_2 it is possible to influence the quality of product without sacrificing amount of product. It is also noticeable that there is some quadratic influence of HS_2^2 with respect to %Fe_FAR and %P_FAR.



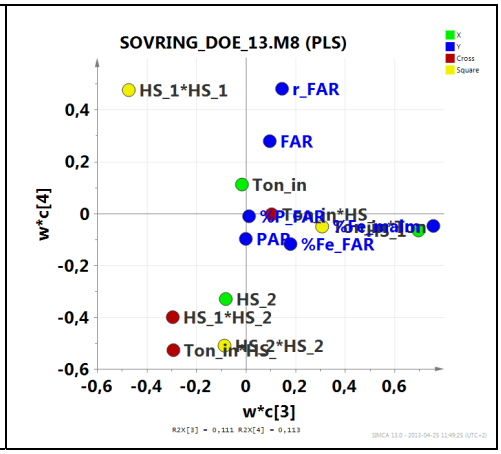


Figure 18.5: (left) PLS $w*c_1/w*c_2$ weight plot. Ton_In is the most influential factor for PAR and FAR. The model indicates that by increasing Ton_In the amounts of PAR and FAR grow. There is a significant quadratic influence of HS_2 on the quality responses %Fe_FAR and %P_FAR. This non-linear dependence is better interpreted in a response contour plot, or a response surface plot. Figure 18.6: (right) PLS $w*c_3/w*c_4$ weight plot.

In order to better assess the importance of the higher-order model terms (cross-terms and square-terms), we decided to inspect response contour plots. Figures 18.7 – 18.10 show response contour plots of the important responses PAR, FAR, %P_FAR, and %Fe_FAR, respectively. These plots were produced by using Ton_In and HS_2 as axes, and by setting

HS_1 at its high level. (Setting HS_1 high was found beneficial using plots of regression coefficients).

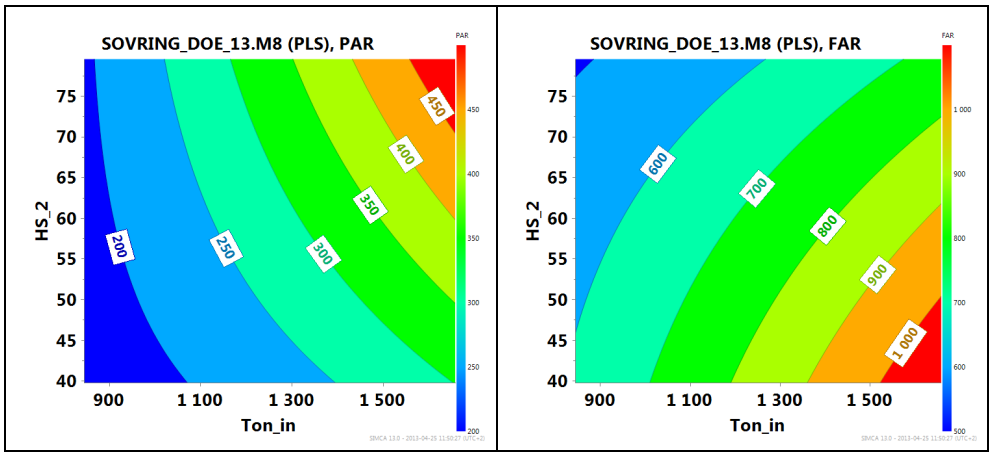


Figure 18.7: (left) Response contour plot of PAR, where the influences of Ton_In and HS_2 are seen. Figure 18:8: (right) Response contour plot of FAR.

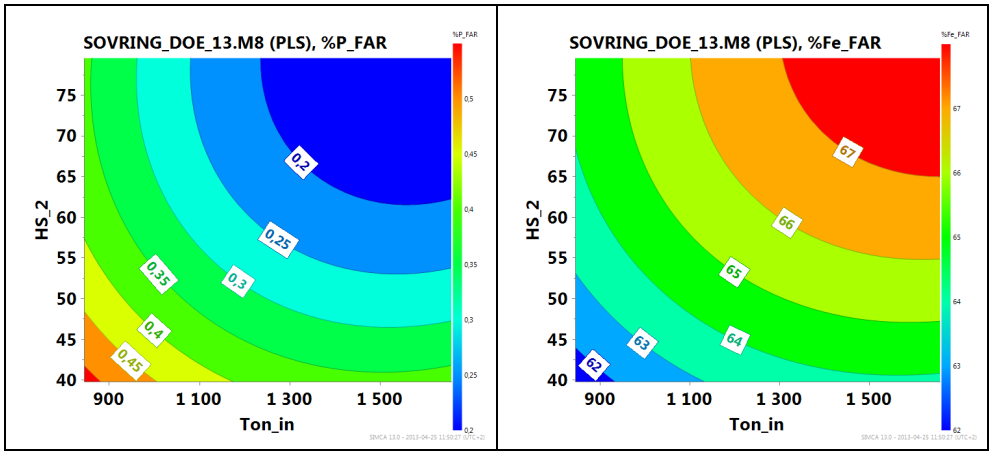


Figure 18.9: (left) Response contour plot of %P_FAR. Figure 18.10: (right) Response contour plot of %Fe_FAR.

The evaluation of the four response contour plots represents a classical example of how the goals of many responses rarely harmonize completely. For three responses, PAR, %P_FAR, and %Fe_FAR, the model interpretation suggests that the upper right-hand corner (high HS_1, high HS_2, and high Ton_In) represents the best operating conditions. At the same time, however, it appears optimal for the response FAR to run the process according to conditions represented by the lower right-hand corner. Staying somewhere along the right-hand edge thus seems a reasonable compromise.

The SweetSpot plot is a practical way to overview optimization results when several response contour plots have to be considered simultaneously. Figure 18.11 shows such a plot for the SOVRING dataset, which was created using the MODDE software. The SweetSpot plot can be thought of as a superimposition of several response contour plots, i.e., an overlay which is color coded according to how well the specifications on the responses are met. If the plot contains a green area, the investigator knows there is a SweetSpot.

We note that the originators of the SOVRING dataset did not provide numerical demands on the response variables; they only indicated whether a response variable should be maximized or minimized. For the sake of illustration we have defined fictitious, but still reasonable, specifications, given the knowledge about which responses should be high and low. Figure 18.11 shows that a SweetSpot region does indeed exist for the SOVRING process.

A limitation of the SweetSpot plot is that it does not display the risk of failure to comply with the response specification. However, by using additional QBD-directed technology inside MODDE, it is possible to evaluate such a risk of failure. Figure 18.12 shows a probability contour plot showing the risk of failure for the SOVRING process. The contour levels indicate how many failures there are in one million attempts. The green area corresponds to an acceptably low risk, 0.1%, i.e., 1000 failures in one million trials. The contour line indicating 50% risk of failure (500000 failures in one million trials) is very similar to the shape of the SweetSpot, which is according to expectation.

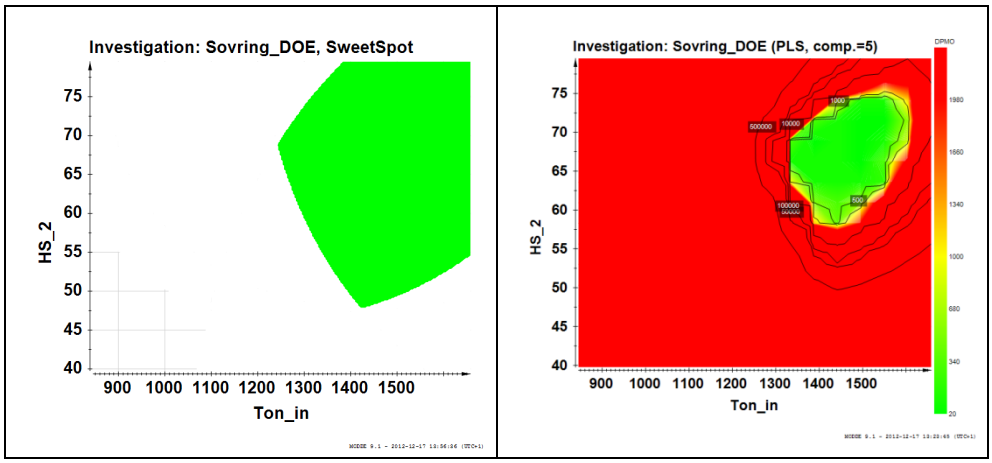


Figure 18.11: (left) SweetSpot plot of the Sovring example, suggesting the SweetSpot to be located in the upper right-hand corner. Figure 18.12: (right) A probability contour plot showing the risk of failure. The green area corresponds to an acceptably low risk, 0.1%, of failure. The 50% risk of failure area is very similar to the shape of the SweetSpot seen in Figure 18.11.

18.3.3 Summary of application

The SOVRING example shows the utility of an optimization design in process industry. Two main types of responses were distinguished: responses related to *amount of product* (i.e., throughput) and those related to *quality of product.* By increasing the load (Ton_In) to the mineral sorting plant, higher production of FAR and PAR resulted. By regulating the speed of the second magnetic separator (HS_2), it was possible to maintain a balance between the two quality responses %P_FAR and %Fe_FAR.

In summary, the net gain of using DOE in the SOVRING example was improved process understanding coupled with better product quality. Sometimes, however, DOE cannot be applied at all, or, at least, not to the extent that is desired. Under such circumstances, the PAT approach might be a viable alternative/complement to DOE/QBD. Section 18.4 is devoted to introducing PAT using the perspective of the pharmaceutical industry.

18.4 PAT: Goals, components, pre-requisites and levels

18.4.1 Goals of PAT according to FDA

The FDA has recognized that significant regulatory barriers inhibit adoption of state of the art manufacturing practices within the pharmaceutical industry. A new risk based approach for Current Good Manufacturing Practices (CGMPs) looks to modernize the regulation of pharmaceutical manufacturing with the goal of enhanced product quality, and allow for continuous improvement leading to lower production costs. This initiative is built on the premise that if manufacturers demonstrate they understand their processes they will be at less risk of producing bad product, will be given the freedom to implement improvements within the boundaries of their knowledge without the need for regulatory review and will become a low priority for inspections. The FDA defines process understanding as the identification of critical sources of variability, management of this variability by the manufacturing process and ability to accurately and reliably predict quality attributes.

The following is a quote from the FDA:

"The goal of PAT is to understand and control the manufacturing process, which is consistent with our current drug quality system: quality cannot be tested into products; it should be built-in or should be by design.

Process Analytical Technology is:

- *a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality.*

It is important to note that the term analytical in PAT is viewed broadly to include chemical, physical, microbiological, mathematical, and risk analysis conducted in an integrated manner.

Process Analytical Technology tools:

There are many current and new tools available that enable scientific, risk-managed pharmaceutical development, manufacturing, and quality assurance. These tools, when used within a system can provide effective and efficient means for acquiring information to facilitate process understanding, developing risk-mitigation strategies, achieving continuous improvement, and sharing information and knowledge. In the PAT framework, these tools can be categorized as:

- *Multivariate data acquisition and analysis tools*
- *Modern process analyzers or process analytical chemistry tools*
- *Process and endpoint monitoring and control tools*
- *Continuous improvement and knowledge management tools*

An appropriate combination of some, or all, of these tools may be applicable to a single-unit operation, or to an entire manufacturing process and its quality assurance.

A desired goal of the PAT framework is to design and develop processes that can consistently ensure a predefined quality at the end of the manufacturing process. Such procedures would be consistent with the basic tenet of quality by design and could reduce risks to quality and regulatory concerns while improving efficiency. Gains in quality, safety and/or efficiency will vary depending on the product and are likely to come from:

- *Reducing production cycle times by using on-, in-, and/or at-line measurements and controls.*
- *Preventing rejects, scrap, and re-processing.*
- *Considering the possibility of real time release.*
- *Increasing automation to improve operator safety and reduce human error.*
- *Facilitating continuous processing to improve efficiency and manage variability*
  - ** using small-scale equipment (to eliminate certain scale-up issues) and dedicated manufacturing facilities.*
  - ** improving energy and material use and increasing capacity."*

18.4.2 Results of PAT

PAT not only provides more insight into raw material quality variances, but allows users to combine spectral and wet chemistry data, and to model the relations between raw material and final product quality. It also allows users to address the complexities of batch monitoring, including:

- Time variance;
- Mixed data, including initial conditions, process data and final product quality data;

- The many unit operations, including granulation, drying, compression, and coating, that are responsible for final product quality.

We will take a hierarchical approach where online processing steps are first modeled and monitored, then the complete batch, including interactions between raw materials and processing steps, is modeled and correlated to quality. The resulting trend lines:

- Summarize normal variation in process parameters;
- Track the status and evolution of the batch;
- Allow users to distinguish between in- and out-of-control operations via control limits on summary variables.

By themselves the multivariate summary statistics are abstract. Contribution plots are therefore provided to depict which variables are responsible for a change in the MVDA space. Such plots connect these multivariate metrics to process parameters to diagnose and remove process faults.

18.4.3 Pre-requisites of PAT

In order to get PAT on a more practical and operational level, we can list a number of pre-requisites:

- *Infrastructure*. Automated data acquisition systems, databases, networks, and synchronization procedures must be in place. The greatest hurdle involved in almost any analysis is generation, integration and organization of data. This is particularly true for the pharmaceutical industry where data are often stored in vast warehouses but rarely, if ever, retrieved and used. Past regulatory environments did not provide incentives for analysis of manufacturing processes because implementing improvements required re-validation and the current condition of pharmaceutical data infrastructures reflects this. As a result, large efforts are required to assemble meaningful datasets. This challenge is further complicated given that laboratory and production data are scattered in various disconnected databases. Examples of these databases include Laboratory Information Management Systems (LIMS), Manufacturing Execution Systems (MES), Enterprise Resource Planning systems (ERP) as well as Supervisory Control and Data Acquisition systems (SCADA) and process historian databases. Product quality is influenced by all stages of production including variability of the raw materials. Developing process understanding of a finished product can only be realized through uncovering the cumulative influence of all processing steps and their interactions. Integrating, synchronizing and aligning data from all relevant sources is therefore a pre-requisite before analysis can begin.
- *Multivariate characterization*. Adequate and informative data must be measured on all steps and ingredients of the process.
- *Multivariate evaluation of all data*. All data should be analyzed together. The data analysis should not focus on variable selection, should not be univariate in nature, and should not involve methods with many adjustable parameters which are prone to overfit. The data analysis phase should entail simple, transparent, informative, and reversible projection models.
- *Data and information integration and communication*. All data flows and data bases should be integrated onto one common platform. This facilitates use of data, visualization of data, and communication of results.
- *DOE*. A suitable use of DOE combined with some of the steps above can augment the analysis and help to ensure that critical system parameters are varied together in a simultaneous and informationally optimal manner.

18.4.4 Four levels of PAT

From a chemometrics point of view, the PAT framework consists of four main steps (levels 1 – 4):

- Level 1 – Multivariate off-line calibration
- Level 2 – At-line multivariate PAT
- Level 3 – Multivariate on-line PAT
- Level 4 – Multivariate on-line PAT of entire process

Level 1 – Multivariate off-line calibration. At the first level, off-line chemical analyses in laboratories are to a large extent substituted by at-line analysis using a combination of rapid measurements (spectroscopy, fast chromatography, images, and sensor arrays) and multivariate calibration. The latter is necessary to convert PAT data into traditional representations, e.g. concentrations, disintegration rates, and so on. Hence, the main objective is to use process analytical data to provide fast and accurate alternatives to time-consuming and laborious laboratory measurements.

Typical applications may include

- The determination of the active pharmaceutical ingredient (API) concentration in the final product or in intermediates (Figure 18.13).
- The determination of concentrations of impurities in the final product as well as in important intermediates.
- The determination of humidity in samples after drying.
- The determination of different crystal forms in API samples, tablets, etc.

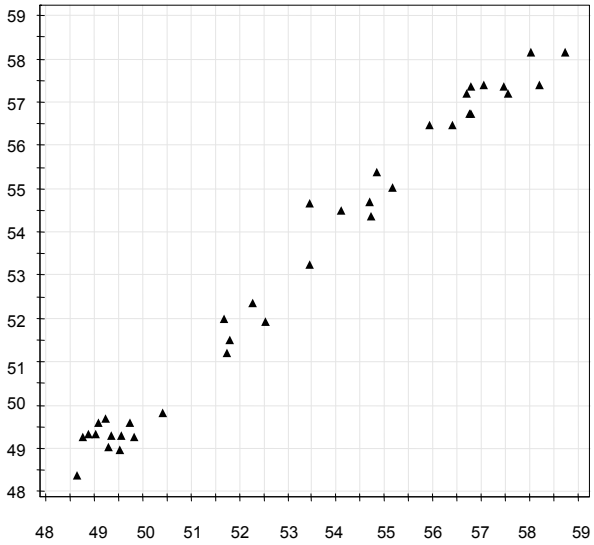


Figure 18.13: Relationship between observed percentage of active substance (abscissa) and predicted percentage of active substance (ordinate) in a pharmaceutical production situation.

Level 2 – At-line multivariate PAT. PAT level 2 includes PAT level 1. At this level, the at-line PAT data are used directly without conversion back to the traditional numerical representation. The objective is often a classification (Figures 18.14 and 18.15) of the intermediate or product as being within specification or not with the same additional objectives as PAT level 1, i.e. reduced waiting time and enhanced accuracy.

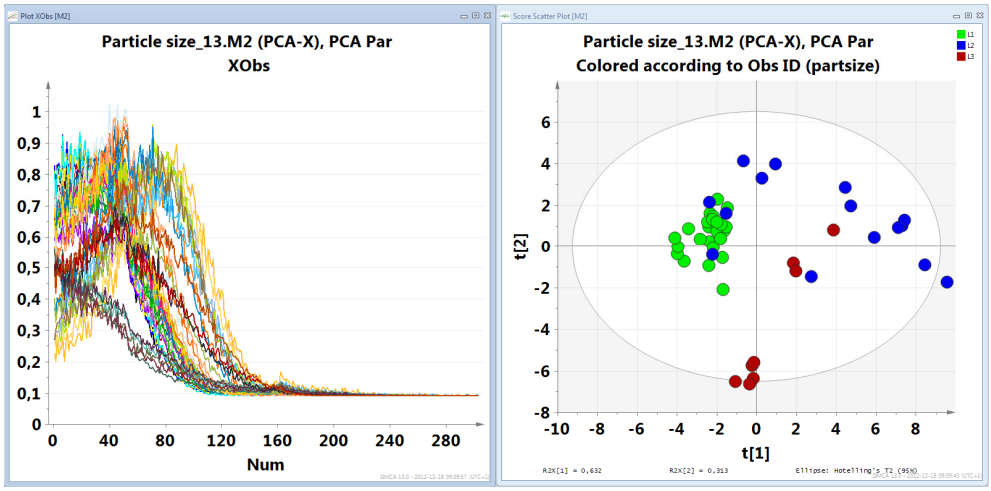


Figure 18.14: (left) An illustration of raw material characterization based on particle size distribution measurements. The data come from a pharmaceutical company and relate to all incoming batches during approximately two years of production. There are in total 300+ variables corresponding to particle size intervals ranging from 2.5 – 150 µ. Figure 18.15: (right) PCA score plot from the analysis of the particle size distribution curves. There are three vendors of excipient. The plot is color-coded according to the information on the vendors. The score plot reveals that supplier L1 provides the most homogeneous and consistent starting material with little quality variation over time.

Level 3 – Multivariate on-line PAT. Here, data are measured also during the batch evolution of each production step – online MSPC and BSPC (see also Chapters 15 and 16). This level includes levels 1 and 2. The objective at this level is real-time process monitoring, where all available data are used to indicate whether the batch process is running normally or not, and whether the batch process has reached normal completion (Figure 18.16). The latter is related to end-point detection, where online PAT data are used to determine the degree of completion in, e.g. the synthesis of the API, a granulator, a dryer, or a tablet coating step.

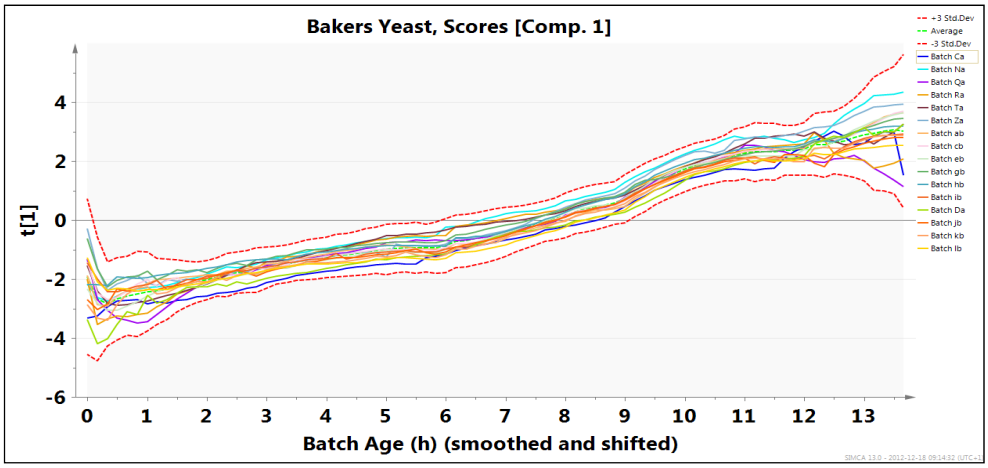


Figure 18.16: On-line PAT: Data are measured during the batch evolution of each production step - online BSPC (Batch Statistical Process Control). Batch progress is conveniently monitored in terms of a BSPC control chart (see Chapters 15 and 16 for deeper discussions of control chart technologies).

Level 4 – Multivariate on-line PAT for the entire process. The final level, which includes levels 1 – 3, implies integrating data from all steps and raw materials together for a total overview of the process, a total overall process or production line fingerprint (Figure 18.17). This integrated knowledge is leveraged for development of high resolution understanding of process variation and its influence on product quality. The objectives include use of process understanding to drive process improvement and knowing in real-time whether or not the final product is within specification, i.e. can be shipped. This concept of parametric release or real time release testing is one major benefit of PAT as outlined by the FDA. The logic is

that if the process has remained within specification at all times during all steps, and if all raw materials were within specifications, then the final product must also be within specification. This reasoning is of course based on the assumption that adequate data have been measured on all parts of the process with sufficient frequency. This demands proper and robust validation schemes and frequent monitoring of all measuring devices.

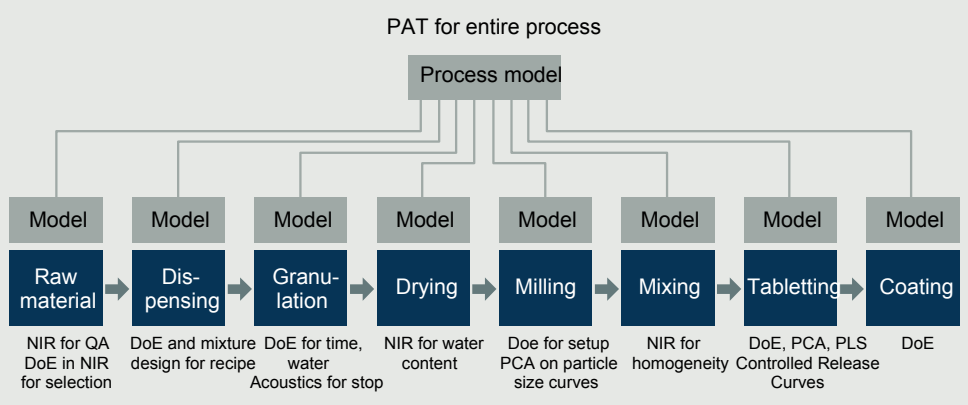


Figure 18.17: Overall on-line PAT: Combining information from all unit operations and raw materials for a complete overview of the entire process. Involves levels 1-3.

18.5 A PAT example: Binary Powder II

18.5.1 Background to example and objectives

Diffuse reflectance NIR spectroscopy is a common method for analyzing powders. It is used for both qualitative and quantitative analysis of powder mixing processes within the pharmaceutical industry. A complicating feature of such studies is that a powder mixture may be homogeneous with respect to its light absorption properties but not with regard to light scattering.

In diffuse reflectance NIR, the spectrum of a powder sample is affected by both the concentration of the excipients and the physical properties of the powder. The measured reflectance is a combination of absorption, refraction and scattering of the incident light. Light scattering properties are influenced by (i) particle size and shape, (ii) powder distribution and packaging, and (iii) chemical composition of the powder.

The current example deals with mixing of two powders with dissimilar particle size, a fine powder with particle size less than 200 μm and a coarse powder with particle size larger than 300 μm. Such a marked difference in particle size may make the system susceptible to segregation and hence real-time monitoring of the mixing process is particularly relevant. Access to this dataset was kindly granted by Dr Ola Berntsson of AstraZeneca, Södertälje [Berntsson, *et al.*, 2000; Berntsson, 2001].

The objectives of the investigation are:

(i) to build a multivariate calibration model for predicting % coarse powder
(ii) to apply the calibration model for PSV of the powder blending process
(iii) to know when mixing is complete.

The example illustrates how PAT can be applied to an important unit operation in pharmaceutical manufacturing.

18.5.2 Primary dataset

The primary dataset consists of 11 batches of powders, prepared and analyzed off-line, with a coarse powder range of 0-100%. The 11 samples were prepared in separate bottles and mixing was achieved by tumbling the bottles. After mixing, NIR spectra were acquired from 100 sub-samples of the powder in each bottle. The bottles were remixed between each

spectrum so that all 100 spectra collected represent different aliquots of the same powder sample. The spectra were obtained as log (1/R) in the 1080-2025 nm (9242 – 4937 cm^{-1}) range, using a nominal resolution of 32 cm^{-1}, giving a total of 280 X-variables and 11*100 = 1100 observations.

The primary dataset was split into two parts, one for model training and the second for model testing. All observations with 0, 20, 40, 60, 80 and 100% coarse powder (600 observations) constitute the training set while the test set is made up of the 10, 30, 50, 70 and 90% samples (500 observations).

18.5.3 Monitoring dataset

A vertical cone mixer was used which is appropriate for materials that are liable to segregation (Figure 18.18). An orbiting transport screw creates a convection mixing mechanism. NIR spectra were acquired by inserting a fiber-optic probe through an interface in the cone wall. In the monitoring experiment, equal amounts of the coarse and fine powders were loaded with the coarse powder on top. A total of 6000 spectra were monitored with approximately 140 spectra corresponding to a single turn of the rotating convective screw.

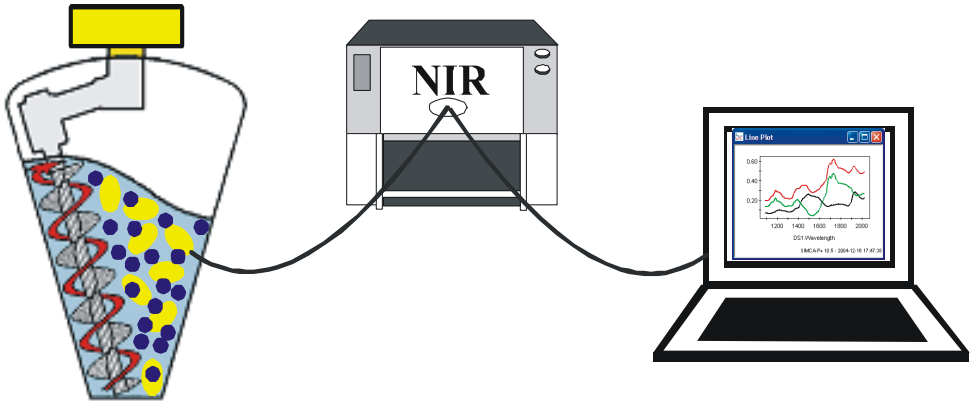


Figure 18.18: Schematic overview of the vertical cone mixer and the fiber-optic probe set-up.

18.5.4 Evaluation of raw data

From an examination of the raw spectra (Figure 18.19), both additive and multiplicative effects are apparent. The higher the percentage of coarse powder the higher the spectral pseudo-absorbances.

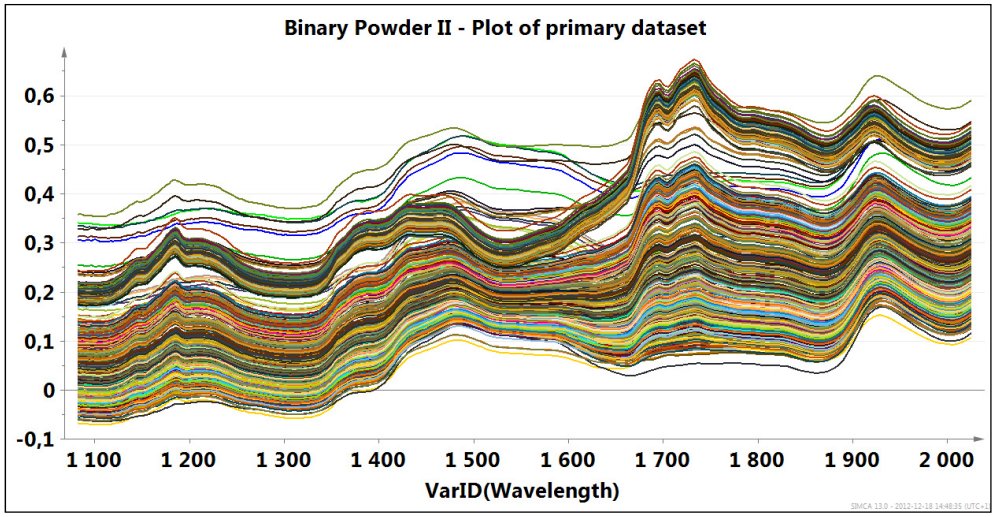


Figure 18.19: The 1100 spectra of the primary dataset.

For interpretation purposes, it is worth looking at the corresponding spectra of the pure mixture components (Figure 18.20). The plot below shows the spectra of the pure powders plus their difference (coarse-fine) spectrum. The pseudo-absorbances are higher for the coarse powder.

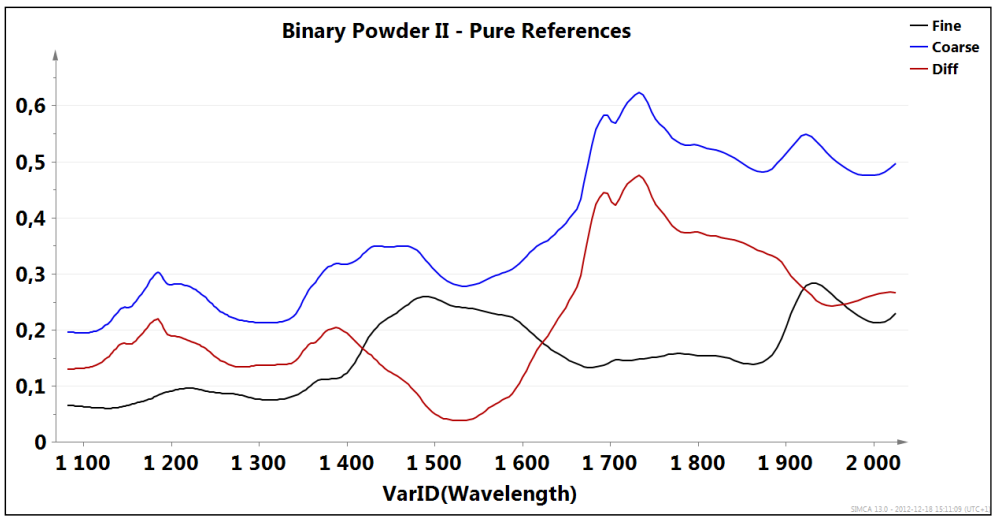


Figure 18.20: Spectra for the "pure" components and the difference spectrum coarse-fine.

18.5.5 PCA for data overview

It is good practice to get an overview of the spectral data using PCA. Relevant questions include: Does the dataset contain groups? Are there any outliers? Is it appropriate to fit a single calibration model or are separate models required? PCA can be performed sample-wise, i.e. for 100 spectra at a time, or across the whole dataset, i.e. looking at all 1100 spectra in one global model. Fitting local models makes it easier to detect outliers and time trends. On the other hand, the global model permits a comparison of the eleven samples, to see whether there are time trends, jumps, discontinuities or other undesirable qualities in the data.

We start by computing local PCA models of the 100 spectra of the sample containing 0% coarse powder, 10% coarse powder and so on. Figure 18.21 shows representative results for three out of the eleven possible local PCA models. All spectra were centered but not scaled so the first principal component in each local model should be dominant. Indeed, the first PC accounts for more than 97% of the variance in all eleven local models. Hence, it is

sufficient to focus only on t_1. Figure 18.21 exemplifies how outliers can be detected using line plots of t_1. All sample models contain some spectral outliers.

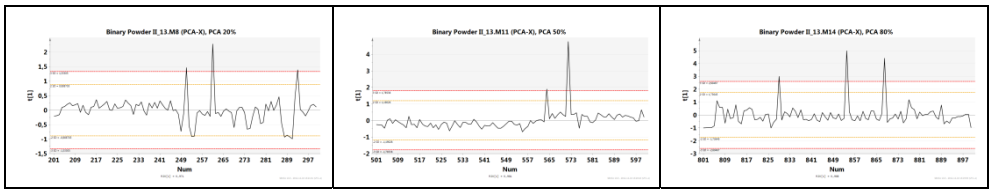


Figure 18.21: Score line plots of t_1 of three sample-wise PCA models, i.e., models for 20%, 50% and 80% coarse powder samples. In each plot the vertical axis corresponds to the score values and the horizontal axis to the number order of the sub-sample spectra.

In the next phase, a global PCA model was fitted to all 1100 spectra to overview the entire dataset and to investigate further the outliers observed previously during the local modeling phase. Two principal components explain 99.9% of the original variation. Note that the first component is far more important (95%) than the second. This needs to be borne in mind when inspecting the score plot shown below (i.e. the horizontal direction is the dominant spectral feature of Figure 18.22).

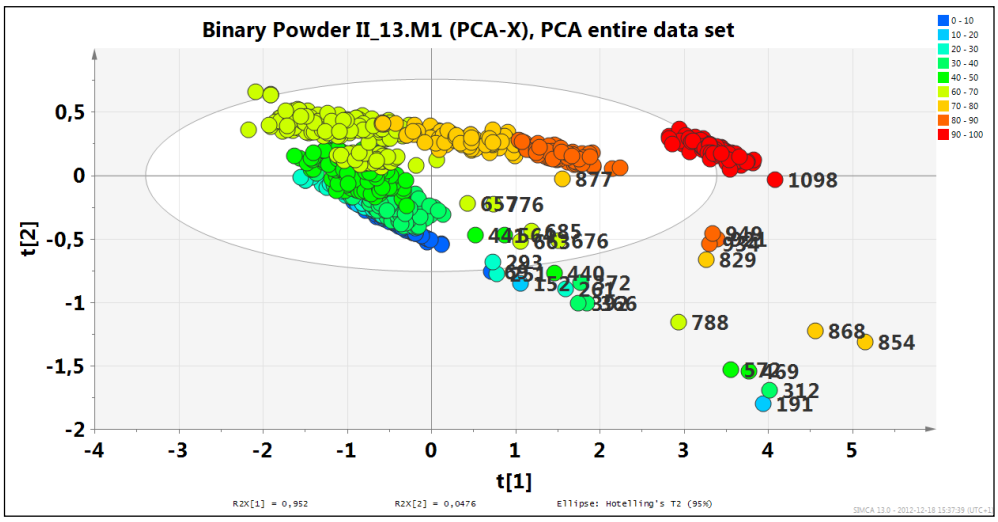


Figure 18.22: Score plot t_1/t_2 showing an overview of the 1100 observations (sub-sample spectra) contained in the primary dataset. The plot is colored according to % coarse powder.

Figure 18.22 highlights a number of outliers, which are far from the other sub-samples in the various batches and hence need to be removed. Most of these were also detected in the local PCA models. We removed the following 29 observations: 69, 152, 191, 251, 261, 293, 312, 366, 372, 392, 440, 441, 469, 564, 572, 657, 663, 676, 685, 776, 788, 829, 854, 868, 877, 921, 934, 949 and 1098. After removing the outliers, the PCA model was refitted. Again, the first component dominates and explains more than 95% of the spectral variation.

In the score plot below (Figure 18.23) there is a clear separation according to % coarse powder. The plot suggests a boomerang shaped structure to the dataset, i.e. a non-linear pattern. Samples with the highest amount of coarse powder are located to the right in the score plot while the second component contrasts low percentage with medium-high percentage coarse powder samples.

There are no new outliers. However, due to the large influence of the samples from the 100% batch, it might become necessary to exclude these observations during the calibration phase to achieve a more balanced soft sensor model.

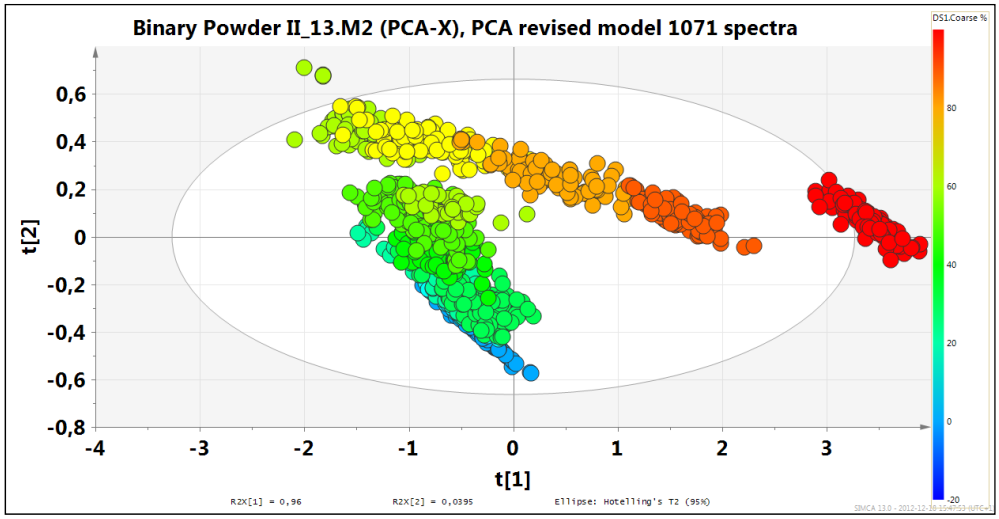


Figure 18.23: Score plot t_1/t_2 of the updated PCA model showing the separation among the various batches depending on the percentage coarse powder. The plot is colored according to % coarse powder.

The loading spectra shown below (Figure 18.24) confirm this interpretation. The first component models baseline differences (recall that high percentage batches have higher baseline offsets). The second component resembles the difference spectrum (coarse powder – fine powder).

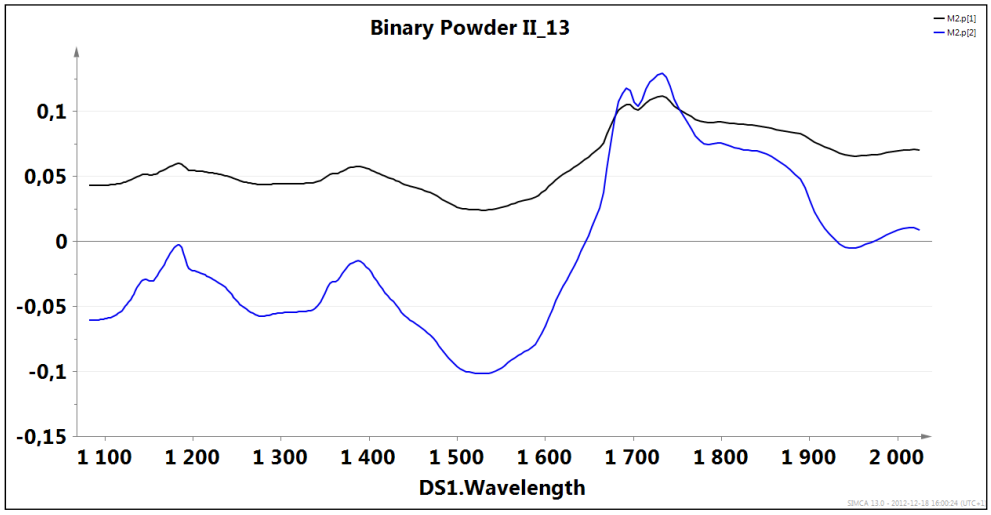


Figure 18.24: Line plot of the loading spectra of the first two principal components.

18.5.6 PLS model based on the training set

A PLS model with 10 components was fitted to the training set (containing 584 – not 600 – observations following outlier removal). The X-scores of the first two components are plotted in Figure 18.25. The grouped and curved pattern observed previously (cf Figure 18.23) is preserved. The strong non-linear structure is clearly seen in the t_1/u_1 score plot shown in Figure 18.26.

This non-linearity can be explained by the large difference in scattering properties between the two powders due to their differential particle sizes. One way to try to reduce the degree of non-linearity is to log-transform the spectral data. This was tried but the results did not change appreciably and the non-linear behavior was still apparent. Hence, in the following, we will focus on the untransformed spectral data. Another option would be to eliminate the samples from the 100% batch as these contribute most to the observed non-linearity.

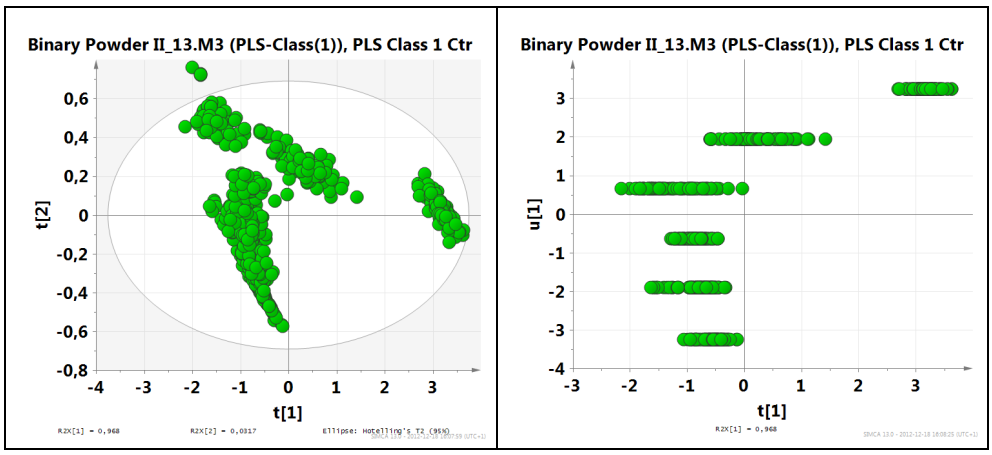


Figure 18.25: (left) PLS X-score plot (t_1/t_2). The plot shows the relationship among the 584 spectra used to calibrate the PLS model. Figure 18.26: (right) PLS X/Y score plot (t_1/u_1) suggesting a strong curvature to the structure between X (NIR data) and Y (% coarse powder).

The loading spectra of the first three components are shown below (Figure 18.27). The first component is related to the high percentage batches and therefore resembles the pure spectrum of the coarse powder. The second loading spectrum mimics the difference spectrum (coarse powder – fine powder). The third loading spectrum is hard to ascribe to any specific chemical or physical phenomenon but captures information in the vicinity of the peaks at 1200 and 1500 nm.

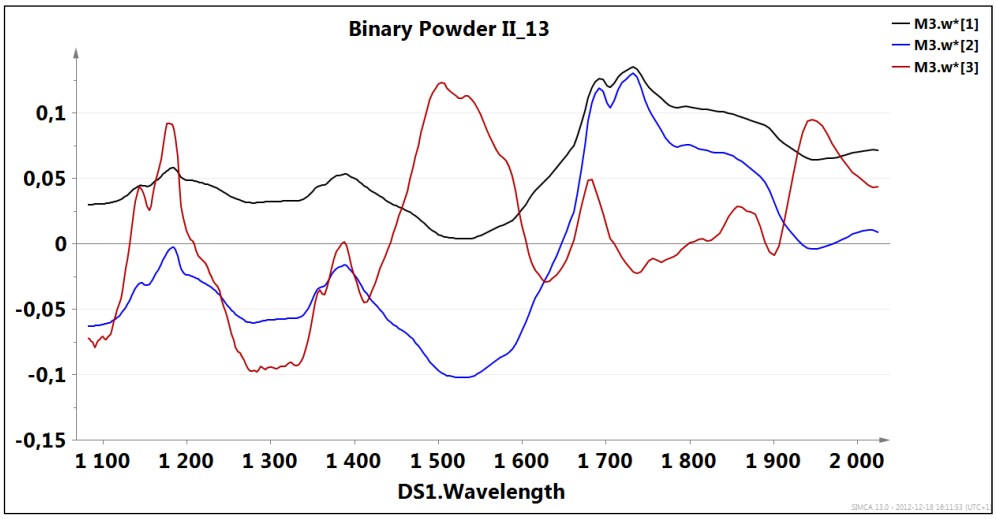


Figure 18.27: Line plot of the PLS loading weights (w) of the first three components.*

Subsequently, the calibration model was applied to the test set (containing 487 – not 500 – observations following outlier removal). It can be seen from the t_1/t_2 plot (Figure 18.28) that the training set is representative of the prediction set. Figure 18.29 displays the agreement between observed and predicted Y-values.

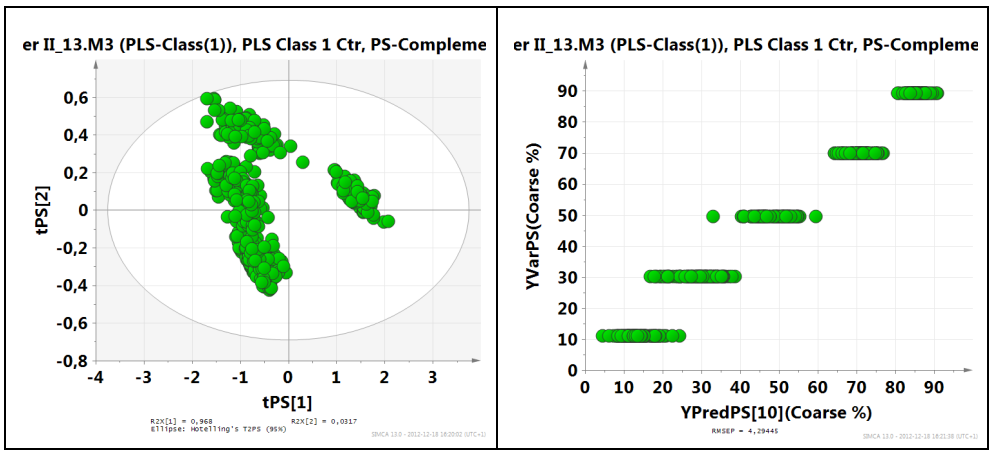


Figure 18.28: (left) Projection of the test set observations onto the t_1/t_2 plane defined by the training set. The training set is representative of the prediction set. Figure 18.29: (right) Relationship between observed and predicted % coarse powder for the test set samples.

In summary, according to the observed vs. predicted plot, the relationship after 10 components is practically linear. Hence, the developed linear PLS model is able to cope with the non-linearity, again demonstrating that the training set is representative of the test set. The RMSEP is relatively low (≈ 4.3 %).

Due to the good predictive power on the test set it was deemed appropriate to apply the multivariate calibration model to the monitoring dataset (see next section).

18.5.7 Application to the monitoring dataset

The next step consists of applying the calibration model to the monitoring dataset. However, prior to that it is worth examining the raw spectra (Figure 18.30). Overall, these spectra look similar to those acquired previously during the modeling phase. The two main differences are that (i) the pseudo-absorbances are generally lower and (ii) the broad peak between 1400 and 1650 nm is more pronounced in the blending measurements.

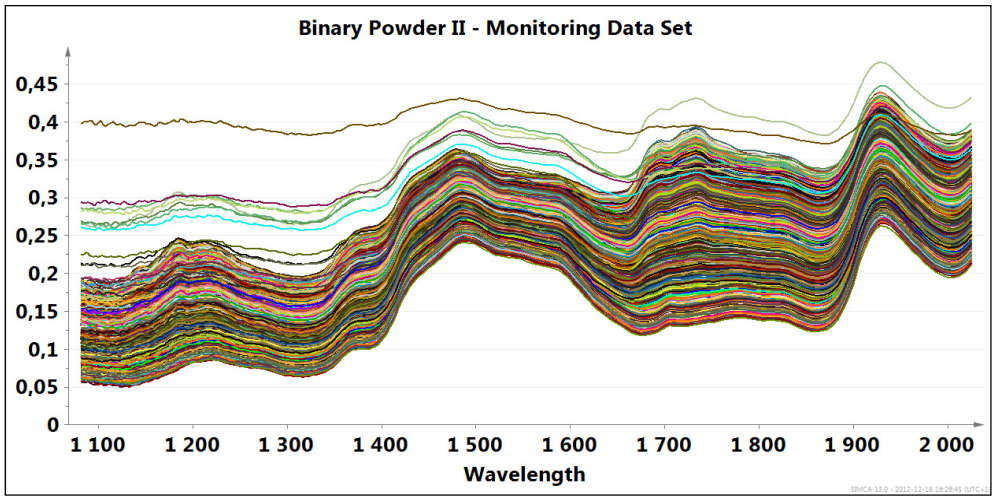


Figure 18.30: The 6000 spectra contained in the monitoring dataset.

A more powerful way of comparing the spectra from the different data sources is by means of the predicted score plot seen in Figure 18.31. The observations from the monitoring dataset are quite different to those from the training set. They are also above Dcrit in the DModX plots across all components (no plots provided). The unavoidable conclusion therefore is that, because the monitoring spectra are too different from the training set, we cannot expect the monitoring task to be of great practical value.

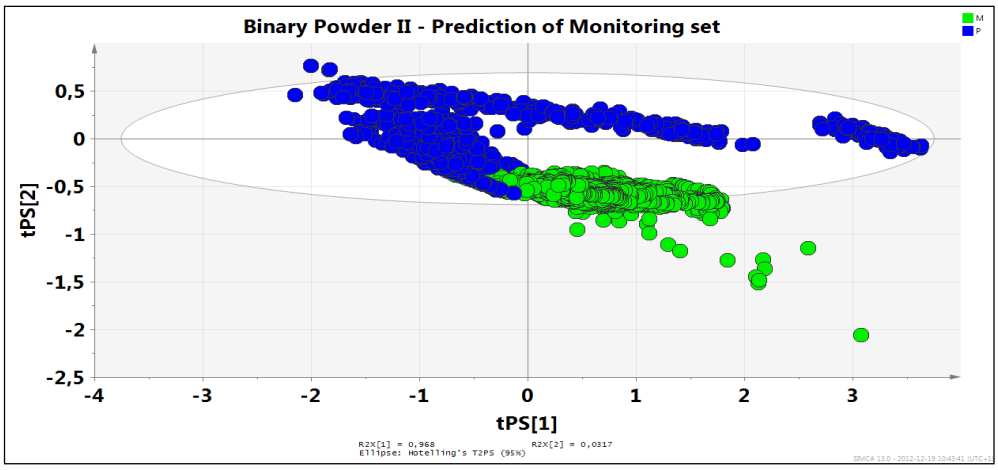


Figure 18.31: Predicted score plot. The monitoring spectra (shown in green) are situated differently compared with the training and test set spectra (show in blue). Hence, we must realize that the calibration model is unrepresentative of the monitoring dataset.

The line plot of the predicted % coarse powder in the mixing process shows great variability at the beginning of the campaign but rapidly stabilizes to a level of about 37-38% (Figure 18.32). The large variation at the beginning is due to the initial complete segregation of the two powders. For each cycle of the orbiting screw the mixing is improved. The line plot contains several spikes, which arise when the rotating equipment passes in front of the probe window. In principle, a spike should be seen every 140th spectrum, however, due to plot resolution restrictions not all spikes are seen. The magnifier functionality in the software provides a better view of when the spikes occur. An alternative would be to create a control chart of the "X-bar s" type using a sub-group size adjusted to the number of scans per cycle of the orbiting screw. This will create a smoother trend curve. Such a plot is illustrated in Section 18.5.10.

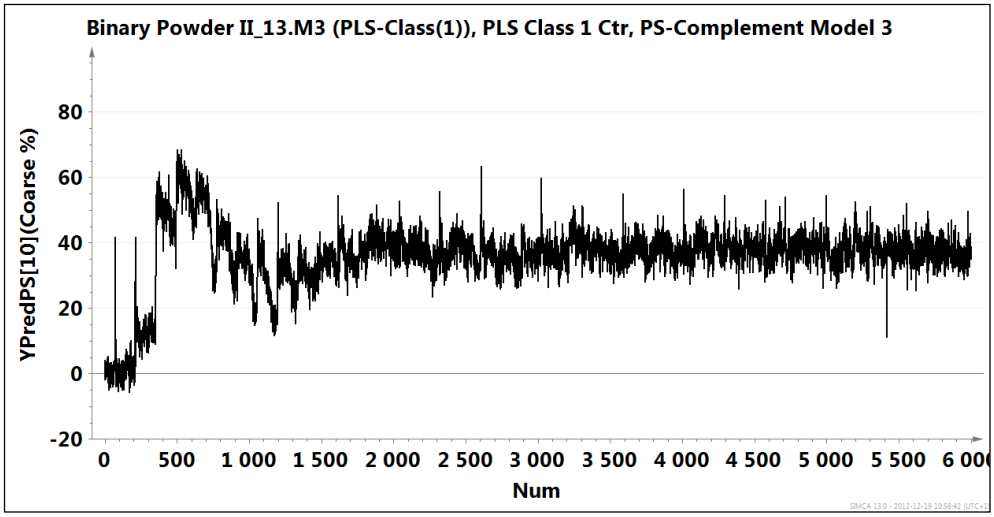


Figure 18.32: Line plot of the predicted % coarse powder.

It can be seen that powder homogeneity is reached about 30-40% into the mixing process (between Num = 1800 to 2400), which is in agreement with the findings reported by the original investigators. However, the predicted level of % coarse powder of 37-38% is not satisfactory. The reason for this deviation from the nominal content of 50% is probably the fact that the training data are not fully representative of the mixing data (recall the mismatch detected in Figure 18.31), at least when working with uncorrected spectral data.

18.5.8 SNV pre-processing for improved predictive power

Because of the failings noted above, we need to sharpen our models using signal (spectral) correction methods. We have tested several approaches and will summarize the results in this section. The Standard Normal Variate (SNV) correction method, which is intended to remove additive and multiplicative baseline shifts, was applied to the spectral data, the effect of which is displayed in Figures 18.33 and 18.34.

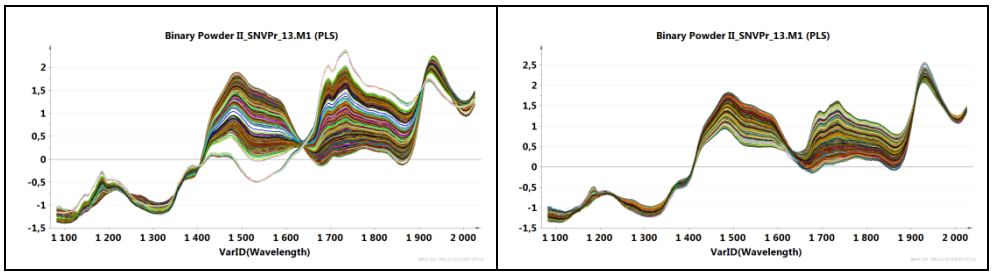


Figure 18.33: (left) Effect of SNV on the 584 spectra in the training set. Figure 18.34: (right) Effect of SNV on the 6000 spectra in the monitoring set.

An updated PLS model with 9 components was fitted to the SNV pre-processed spectral data of the training set ($R^2X > R^2Y > Q^2Y > 0.99$; A = 9). The next two plots (Figures 18.35 and 18.36) display the spread of the observations in the t_1/t_2 plane. The distribution in these two plots should be viewed against the spread seen in Figure 18.31. Evidently, the SNV pre-processed spectra are much more coherent and harmonized.

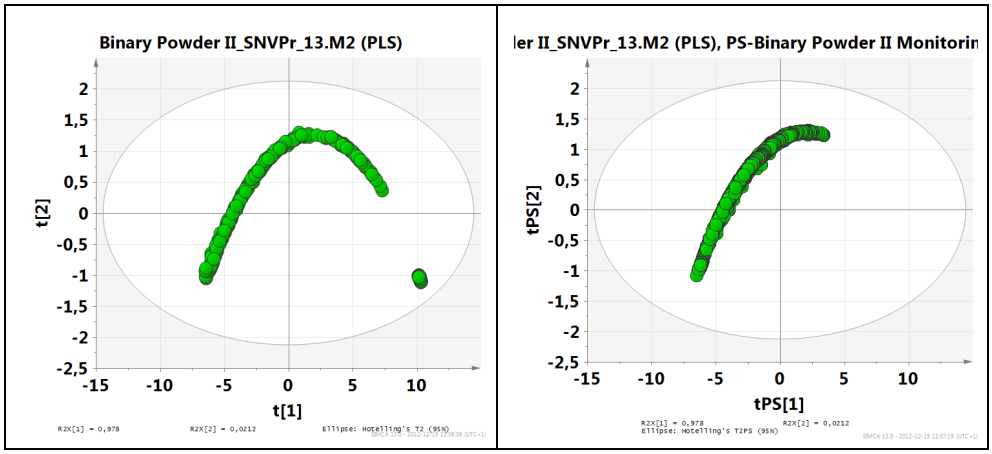


Figure 18.35: (left) First X-score plot of the calibration model based on the SNV-corrected NIR-data. The plot shows the distribution of the 584 training set spectra. Figure 18.36: (right) Predicted score plot corresponding to the foregoing plot, but demonstrating how the 6000 monitoring spectra are distributed.

The prediction of the monitoring set has improved substantially (Figure 18.37). We are now much closer to the nominal value of 50% coarse powder denoting mixture completion. A horizontal line depicting 50% coarse powder is inserted in the line plot.

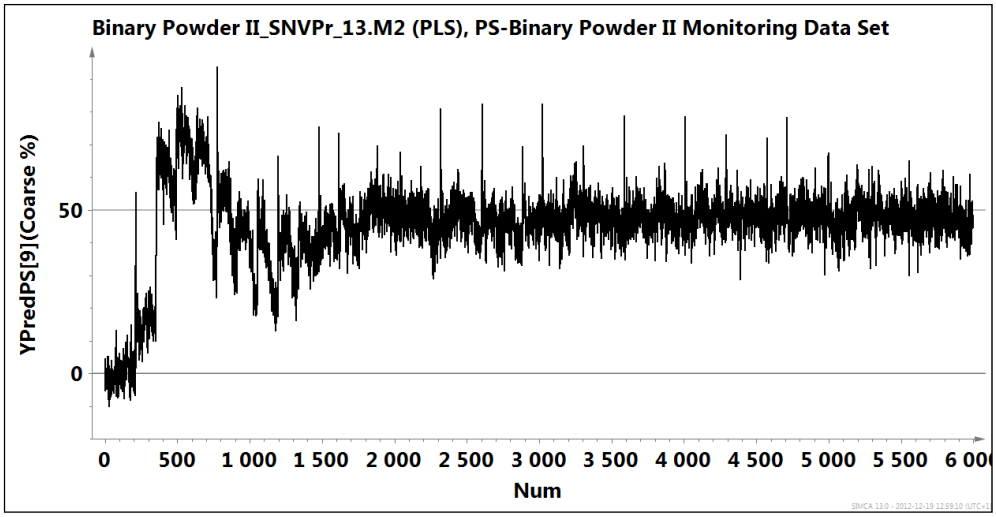
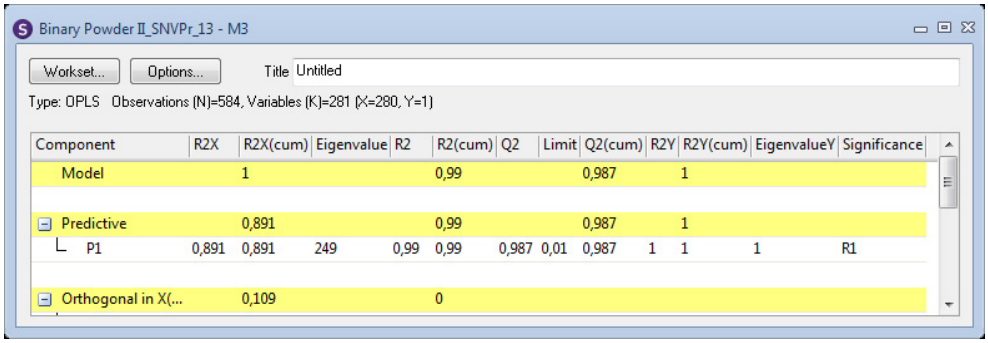


Figure 18.37: Line plot of predicted % coarse powder for the monitoring data set. The plot suggests that a homogeneous mixture is achieved after 30-40 % of the time equivalents.

18.5.9 OPLS for enhanced interpretability

As discussed in Chapter 5, OPLS can be used to re-express a multivariate calibration model in terms of components orthogonal to Y and components correlated to Y. We applied OPLS to the SNV corrected spectral data. This produced an OPLS model consisting of nine components, one predictive and eight orthogonal-in-X. The modeling statistics are displayed in Figure 18.38. OPLS finds that 89% of the spectral variation is relevant for predicting the % of coarse powder.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		1			0,99			0,987		1		
Predictive		0,891			0,99			0,987		1		
P1	0,891	0,891	249	0,99	0,99	0,987	0,01	0,987	1	1	1	R1
Orthogonal in X(...		0,109			0							

Figure 18.38: Summary of fit table for the OPLS model fitted to the SNV corrected data.

One way of comparing the PLS and OPLS models consists of comparing the scaled and centered PLS regression coefficients with the Y-related profile of OPLS. Recall that the basics of the Y-related profile coefficients were discussed in Section 5.4.4. The PLS regression coefficients after nine components are plotted as a line plot in Figure 18.39 and the corresponding OPLS Y-related profile is rendered in Figure 18.40. The OPLS Y-related profile is much smoother and much more suggestive of real spectral features.

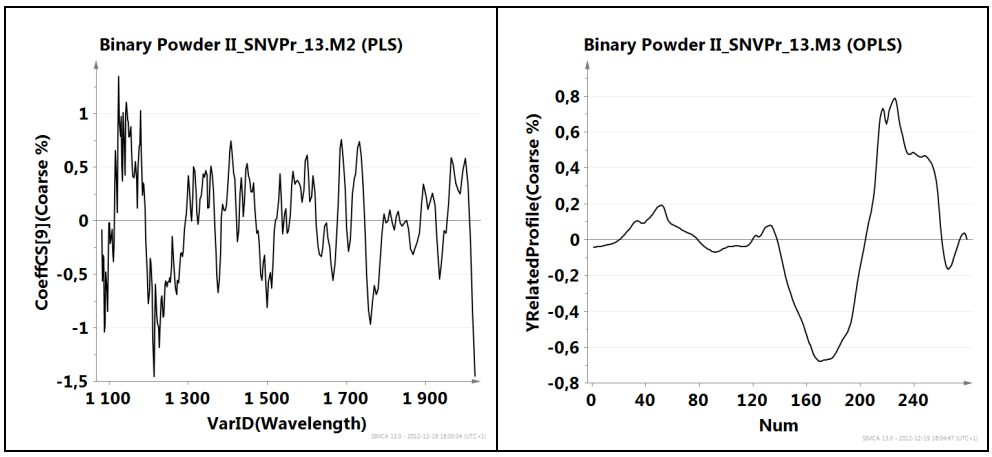


Figure 18.39: (left) PLS regression coefficients after nine components. Figure 18.40: (right) OPLS Y-related profile.

Figure 18.41 provides additional information on the pure "constituents" (powders), which is helpful when interpreting the meaning of the OPLS Y-related profile. Figure 18.41 shows SNV-corrected pure spectra for the fine and coarse powders plus the difference spectrum between them. The OPLS Y-related profile is plotted in the same plot. Note that all profiles are scaled to fit in the same plot. There is a striking similarity between the OPLS Y-related profile and the difference spectrum, particularly at the high-wavelength region end of the spectral window.

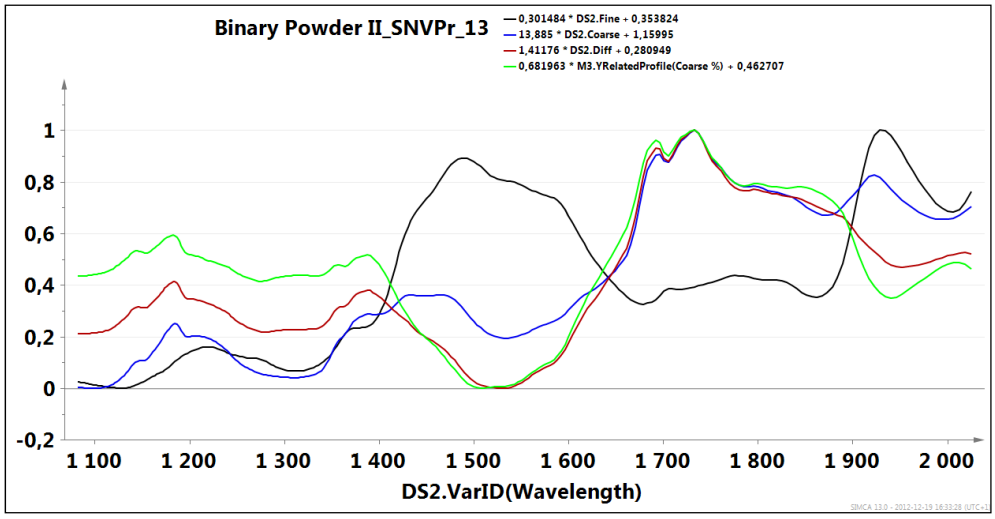


Figure 18.41: Plot of SNV-corrected pure constituent spectra and the OPLS Y-related profile coefficients. Note that all spectra are scaled to fit well in the same plot. The blue line represents the fine powder, the black line the coarse powder, the red line the difference, and the green line the Y-related profile coefficients. There is a striking similarity between the Y-related profile and the difference spectrum.

With OPLS it is easy to visualize the effect of the division into predictive and orthogonal components. The two inner relation (t_1/u_1) score plots in Figures 18.42 and 18.43 reveal that OPLS removes a non-linear structure from the t/u relation.

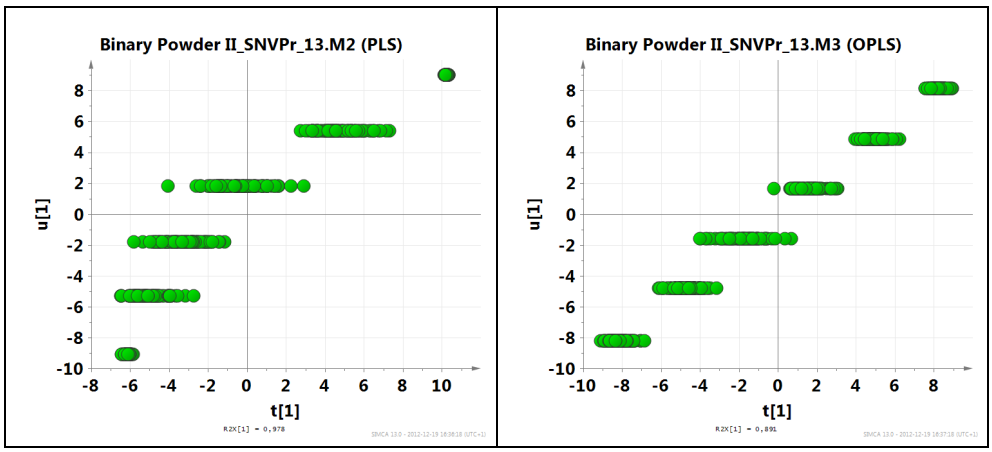


Figure 18.42: (left) PLS t_1/u_1 score plot of the model based on SNV-corrected spectral data. This plot is comparable to Figure 18.26. Figure 18.43: (right) OPLS t_1/u_1 score plot of the model based on SNV-corrected spectral data. This plot is comparable to Figure 18.26.

Another interesting aspect of the OPLS model is that it allows the monitoring data to be plotted as a single control chart of predicted t_1 (Figure 18.44). When inspecting Figure 18.44 it should be noted that the average predicted t_1 score for the 100 test set spectra with 50% coarse powder is -0.29. Hence, we again see that homogeneous mixing occurs within 30-40% of the sampling period.

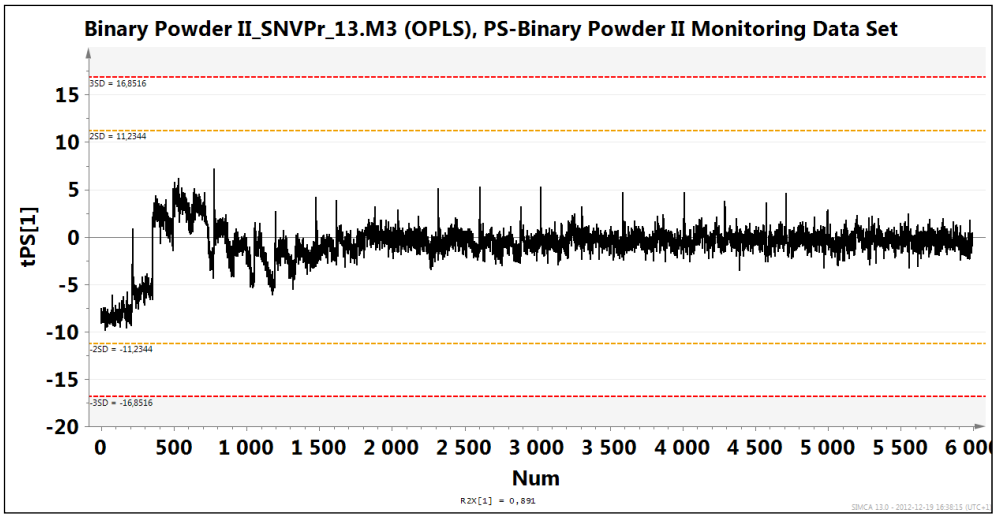


Figure 18.44: Control chart of predicted OPLS t_1-score values of the mixing process.

18.5.10 Discussion of example

This example clearly demonstrates that NIR is an appropriate instrumental technique for modeling the relative amounts of the two powders. This indicates the suitability of NIR/PLS/OPLS in pharmaceutical production unit operations, such as mixing and blending, relevant in PAT.

Berntsson *et al.* have developed a simplified methodology for binary powders, based on using relatively few calibration samples prepared and analyzed off-line [Berntsson, *et al.*, 2000; Berntsson, 2001]. The traditional approach to calibration is to collect one spectrum from each of a number of calibration samples with known concentrations. However, when heterogeneous powder samples are analyzed with reflectance spectroscopy, it is not possible to assign a "true" concentration value to each spectrum. Instead, several spectra must be acquired from different sub-fractions of each sample. The mean of these spectra is then used in the calibration process to represent the nominal content of the sample.

From the results seen above, there is no doubt whatsoever that the methodology developed by Berntsson *et al.* enables quantitative *in situ* monitoring of mixing processes using NIR. NIR captures information about the mixture homogeneity in real time. The authors advocate a high rate of in-line sampling (spectral acquisition) to get a detailed characterization of the variation within the mixture. This method represents a useful tool for examining the influence of various process parameters on the mixing time and mixing quality during the development and scale-up of the process.

In a monitoring situation it is of interest to evaluate critical parameters as a function of time. This is commonly done in multivariate statistical process control (MSPC). Within MSPC, one may chart multivariate parameters, such as scores, Hotelling's T^2, DModX, predicted Y, etc., using the conventional SPC control chart framework (Shewhart, EWMA & CuSum). As an example, the plot below (Figure 18.45) provides a Shewhart chart of the predicted level of % coarse powder in the mixing process monitored in-line and over time. The process operator can easily follow, in real-time, how the mixing process evolves. To facilitate interpretation, the progress of the mixing process is compared with the target value and upper and lower warning and control limits (here corresponding to ±2 and ±3 standard deviations).

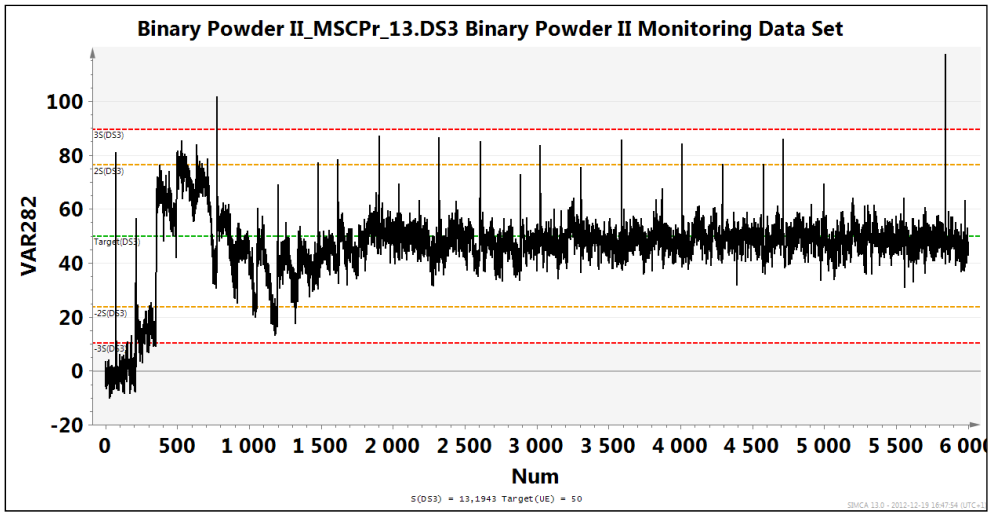


Figure 18.45: Shewhart chart of predicted % coarse powder in the mixing process.

The plot above reveals that the mixing process is complete and stable from around 30-40% into the sampling period. The spikes, which are due to the rotating screw passing in front of the probe window, are clearly seen. One way to get rid of these spikes is to use sub-grouping and co-chart the average Y-predicted value and the surrounding standard deviation inside each sub-group. Such a control chart is shown below (Figure 18.46). When constructing this plot, we used a sub-group size of 14, which means that 10 units on the horizontal scale correspond to one complete turn of the orbiting screw.

The effect of the sub-grouping is a smoothed version of the Y-predicted trend curve. The spikes are now evident in the standard deviation control chart as there is a peak at every 10th point (6, 16, 26, 36, etc.) in the control chart. Alternatively, one may choose to use a larger sub-group size (say 140), but with larger size numbers there is a risk that the inherent information and variation is partially "erased" from the control chart.

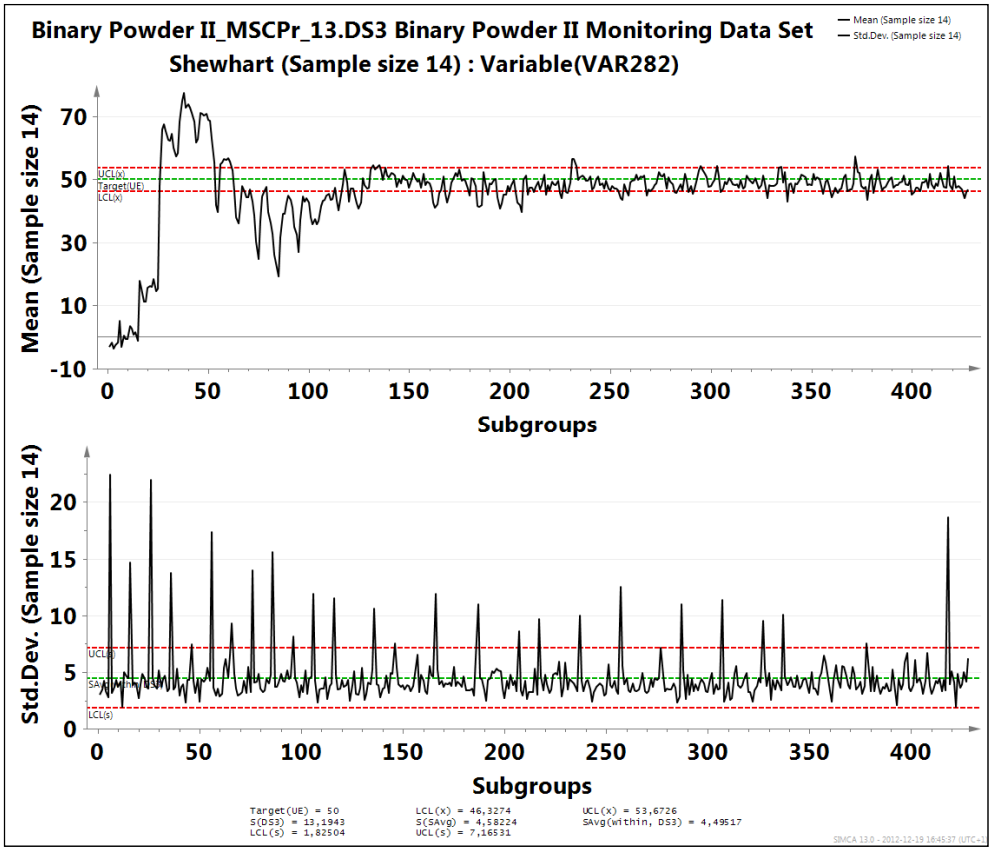


Figure 18.46: Tiled Shewhart charts of average Y-predicted value and the surrounding standard deviation inside each sub-group.

In summary, in the current application, spectral pre-processing is necessary to get a working in-line monitoring model. The combination of SNV and OPLS seems to be particularly powerful. In this study, the external RMSEP is approximately 4.3%. The uncontrolled experimental variation, generated by holding the fiber-optic probe against the powder surface in 100 randomly chosen positions, may inflate the overall RMSEP. By comparing the average predicted percentage coarse powder across the 100 observations of a powder batch with the nominal content, it is estimated that roughly 50% of the overall RMSEP is due to this instrumentation variability.

Thus, we strongly believe the calibration model obtained is useful for in-line monitoring of the mixing process. We note that such a use of mixing process monitoring during routine production is in line with the paradigm of real time parametric release, which focuses on continuous quality assurance and improvements during all production steps, rather than solely on the final step and final product.

18.6 Combining the concepts: The Novartis PAT example

18.6.1 Background

The last dataset to illustrate some of the concepts described in this chapter comes from a feasibility study made together with Novartis in Suffern, NY, USA. We are grateful to James Cheney, John Sheehan, and Fritz Erni for granting us permission to show this example.

In this pharmaceutical production process, the manufactured tablets showed an unacceptable tendency to dissolve too rapidly (dissolution rate less than 90). Process variables and descriptions of the raw materials going into the tablets were available for a period of just over two years. In the present investigation, only process data related to the dryer were

included because the operation of the dryer was suspected of causing the product quality problems.

The process data comprise N_{tot} = 5743 observations with 6 process variables (e.g. temperature, air flow, etc.) + process time for each. These observations are divided into N_{bat} = 313 batches, which in turn are divided into 2 phases (steps). In addition, data for seven raw materials (excipients) are available for each batch, as well as summary data for other process steps (granulation, solvent addition, film coating), giving 298 variables in total. Finally, 21 quality measurements are available for each batch, where the dissolution rate is the most important.

Note that these data reside in different databases and are unaligned and unsynchronized so a key part of the data analysis is to ensure that the batches are comparable and that the right raw material data are aligned with the right batch.

18.6.2 Workflow to synchronize data

In this case, the data have different dimensions and sizes and it is therefore necessary to align them to put them on a comparable footing. Getting the different types of data into a single data structure was made according to the principles of hierarchical and batch-wise analysis of process data outlined in [Wold, *et al.*, 1996; Wold, *et al.*, 1998b], see also Chapters 16 and 19. Figure 18.47 shows an overview of the models' structure.

- Step 1. First, the data of the two phases were arranged in accordance with Figure 16.6. Two separate PLS models were made, one for each phase (step) of the process using local process time as y (Figure 16.6).
- Step 2. Subsequently, the original process variables were chopped up in one piece per batch, transposed, aligned by linear time warping, and used as descriptors of the process dynamics. The resulting matrix with one line per batch was then subjected to a PLS analysis with y = dissolution rate. This gave 2 lower level models (Figure 18.47).
- Step 3. The data for each raw material (one row per batch) were fitted in separate PLS models with y = dissolution rate. This gave an additional 7 lower level models (Figure 18.47).
- Step 4. The scores resulting from the 9 lower level PLS models developed in steps 2 and 3, in total 36 score variables, were used as variables in the "top hierarchical" batch model, again with y = dissolution rate (Figure 18.47).

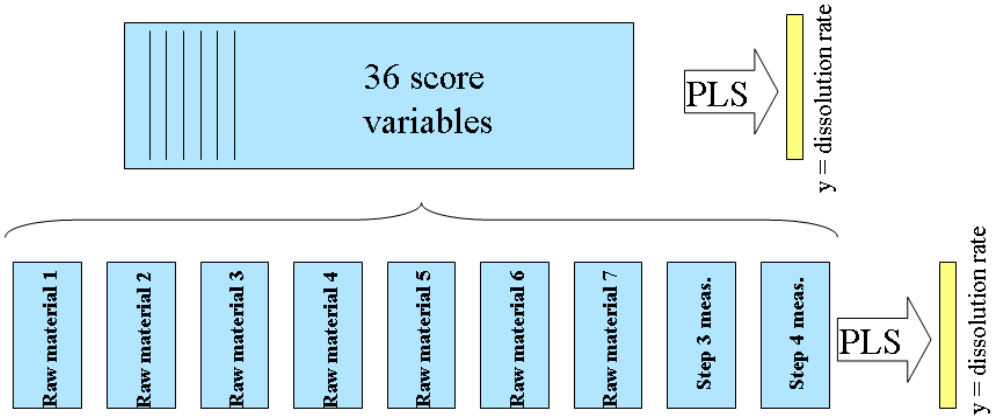


Figure 18.47: Overview of the modeling workflow to get a synchronized data structure. The structure seen is a combination of batch analysis and hierarchical modeling. Each row is one batch.

The result of these three steps was that all the data were combined into a single, consistent and synchronized structure. This allows the PLS estimation of the relationship between y = dissolution rate and X = all data from the process, including raw materials, summarized steps, and the full dynamics of the dryer.

18.6.3 Nine base level batch models

The PLS analysis was, as described in Section 18.6.2 above, divided into two hierarchical levels. At the base level, 9 separate PLS models were built, each with the same y-variable (dissolution rate), while the X-variables were from the two process steps and the 7 raw materials. These 9 base models resulted in between 3 and 6 components each and a total of 36 score vectors. The models had R^2X values of between 0.62 and 0.93 with an average of 0.72. The average R^2Y was 0.15.

18.6.4 Top level batch model

The 36 base level score vectors were then used as X-variables (UV-scaled) in the top level hierarchical PLS model. A two-component PLS model had $R^2X = 0.25$, $R^2Y = 0.44$ and $Q^2Y = 0.39$. The X-scores are plotted in Figure 18.48 and the relationship between observed and predicted dissolution rate is shown in Figure 18.49. The score plot clearly contains clusters and, by coloring by time, it becomes evident that later batches with poor dissolution properties are clustered to the left (negative end of t_1).

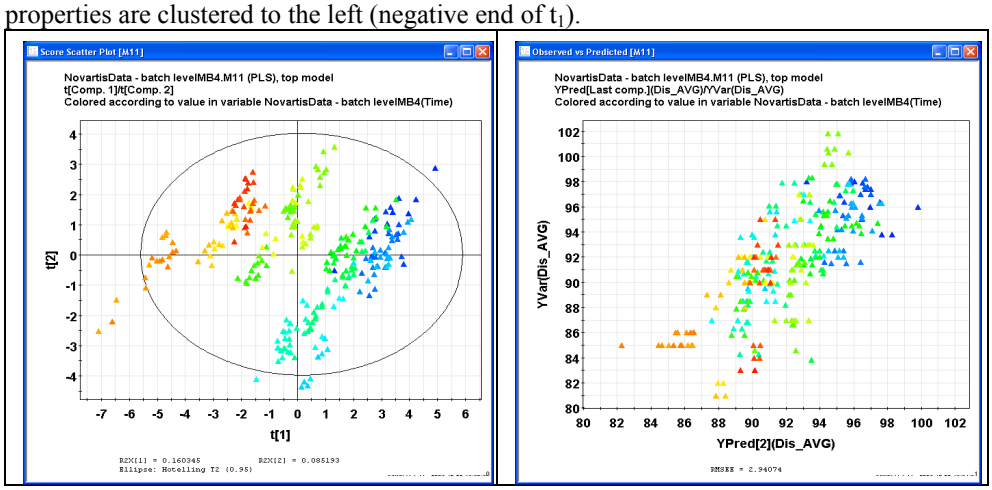


Figure 18.48: (left) Score scatter plot of t_1/t_2. A high proportion of late batches with poor dissolution properties are clustered in the left-hand region of the plot. Figure 18.49: (right) Agreement between measured and predicted dissolution rate for the training set.

The model was validated by a permutation test (Figure 18.50), and by a final model including more complete process data and its real on-line predictions (Figure 18.51).

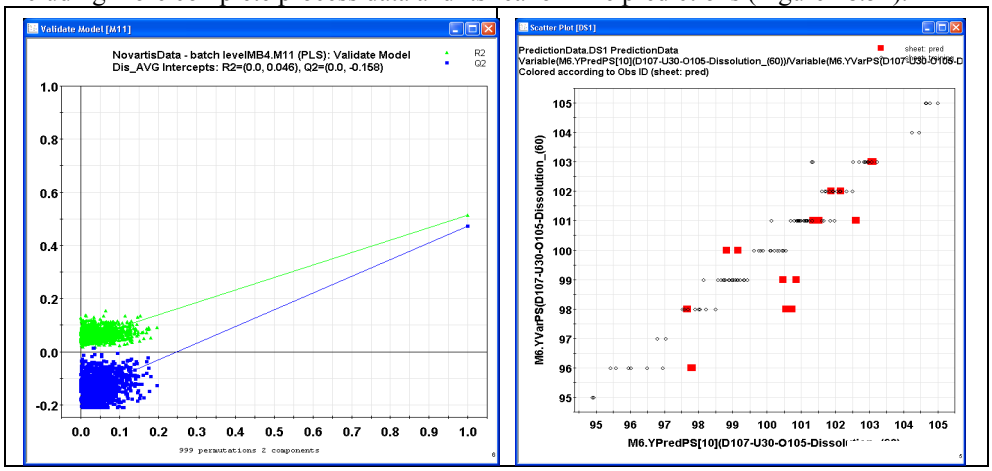


Figure 18.50: (left) Validation plot based on 999 permutations of the y-vector followed by the fitting of a PLS model between X (unperturbed) and the permuted y. The horizontal axis shows the correlation between the original and the permuted y, and the vertical axis the values of R2 (upper line, green) and Q2 (lower line, blue). The plot where all Q^2 values of the permuted y models are below zero is a clear indication that the original model does not happen by coincidence. Figure 18.51: (right) Observed y = dissolution rate on the vertical axis plotted versus the predicted y from the top hierarchical PLS model later developed from more complete process data. Large points are predictions for new batches. $R^2 = 0.82$.

The top model PLS weights of the two components are plotted in Figure 18.52. Several raw materials as well as one phase of the dryer are indicated as important for the changes in y = dissolution rate. Further interpretation is available in [Wold and Kettaneh, 2005].

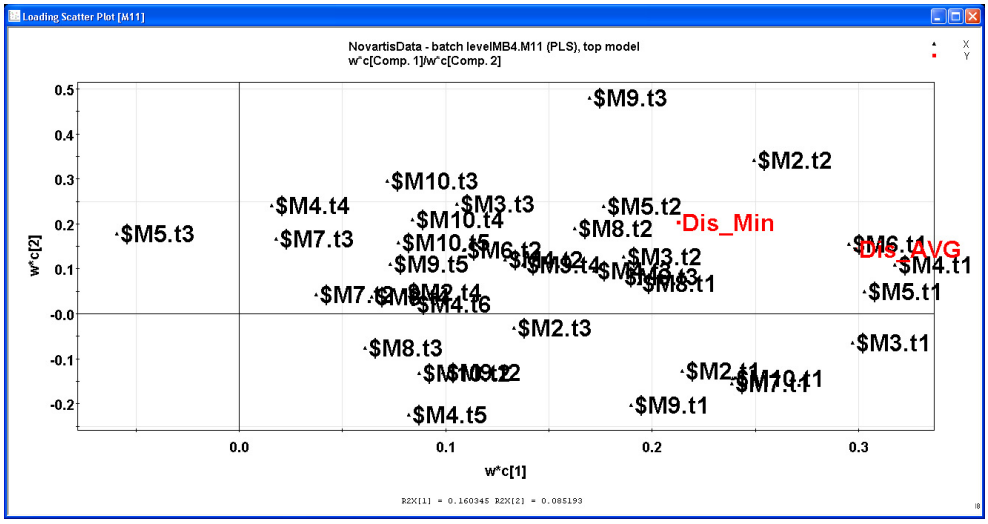


Figure 18.52: PLS-weights (w) for the first (horizontal axis) and second components (vertical axis) of the hierarchical top level PLS model.

18.6.5 Discussion and Epilogue

This hierarchical PLS modeling / process data mining exercise indicated that the process quality variation could be explained. This led to the start of a much larger project looking at all relevant data for both off-line modeling and on-line monitoring.

About a year after the first feasibility study, a system collecting, integrating, aligning, and synchronizing all process and raw material measurements was in place at Novartis, Suffern. A final PLS model was developed from all these data giving a remarkably good (and validated) relationship with dissolution rate (Figure 18.51). This model was put on-line and faithfully predicted dissolution rate for several months (large points in Figure 18.51). This is a remarkable example and demonstrates how good models can be obtained, even when dealing with extremely complex processes, provided that relevant data are available from all parts of the process and that these data are properly aligned.

18.7 Questions for Chapter 18

1. What is PAT?
2. Which are the different levels of PAT?
3. What are their objectives?
4. What is QBD
5. What is a design space?

18.8 Summary and discussion

The PAT concept encompasses a wide range of analytical technologies for in-line, on-line or at-line measurements of pharmaceutical manufacturing processes with the basic objective to determine whether the process and its intermediates and final product are “within specifications”. A successful PAT implementation is characterized by a smooth integration of process data sources (MES, ERP, LIMS, etc.) and instrumentation with computers and pertinent software for data acquisition, data reduction, data analysis, storage of data and

results in databases, and presentation of results including the basis for the final batch record, and when warranted, alarms and diagnostics indicating causes of problems and their severity.

The implementation of PAT in a given process is not automatic and is not achieved overnight. However, with a correct investment in adequate analytical measuring equipment, computer networks, databases, chemometrics and other software, and training of personnel, the return on these investments can be both large and immediate. Waiting time between sampling and analytical results are dramatically reduced or even eliminated, the reliability of results improved, scrap reduced, and process understanding improved.

As mentioned in Section 18.4.4, real time release testing (RTRT) is a major benefit of PAT. To accomplish RTRT, the vast array of model types discussed in this book can be used. These models include multivariate calibration models, surrogate models, design space models and MSPC models. A surrogate model is a multivariate model based on fast process measurements that is replacing traditional and time-consuming measurements. When used in conjunction, such models ultimately provide a higher assurance of product quality.

To quote the ICH Q8 (R2) guidance document, RTRT represents "the ability to evaluate and ensure the quality of in-process and/or final product based on process data". Interestingly, in a presentation at the AAPS Annual Meeting, Washington, October 2011, the acting director of FDA, C.M.V. Moore, discusses a regulatory perspective on RTRT. According to this perspective, RTRT brings about a number of benefits, including (i) increased assurance of quality, (ii) increased manufacturing flexibility and efficiency and (iii) enhanced process understanding. Furthermore, an account is given as to how RTRT relates to QBD. There is no one-to-one relationship between QBD and RTRT, however, as stated by C.M.V. Moore "it would be difficult to justify RTRT without a science and risk based approach".

The obvious extension to using multivariate models and design space models for monitoring and prediction is model predictive control (MPC) of processes [Lauri, *et al.*, 2010; McCready, 2011; Tian, *et al.*, 2011]. Below, we briefly describe an approach to multivariate batch control denoted SIMCA-control, which is based on the work of McCready [McCready, C., 2011]. SIMCA-control is a model predictive control (MPC) method integrated within the multivariate modeling techniques in Umetrics' SIMCA software family (SIMCA-online and SIMCA).

There are two forms of batch control included in SIMCA-control, batch evolution control (BEC) and batch level control (BLC). The objective of BEC is to find the set of manipulated variables (X_{MV}) and their settings that maintain the batch process on a defined batch evolution trajectory (see example in Figure 18.53). An example may include the adjustment of nutrient flows so that a batch evolution model score value is maintained within SPC limits. It is envisioned that BEC is a *regulatory* type of control executed at a relatively high frequency.

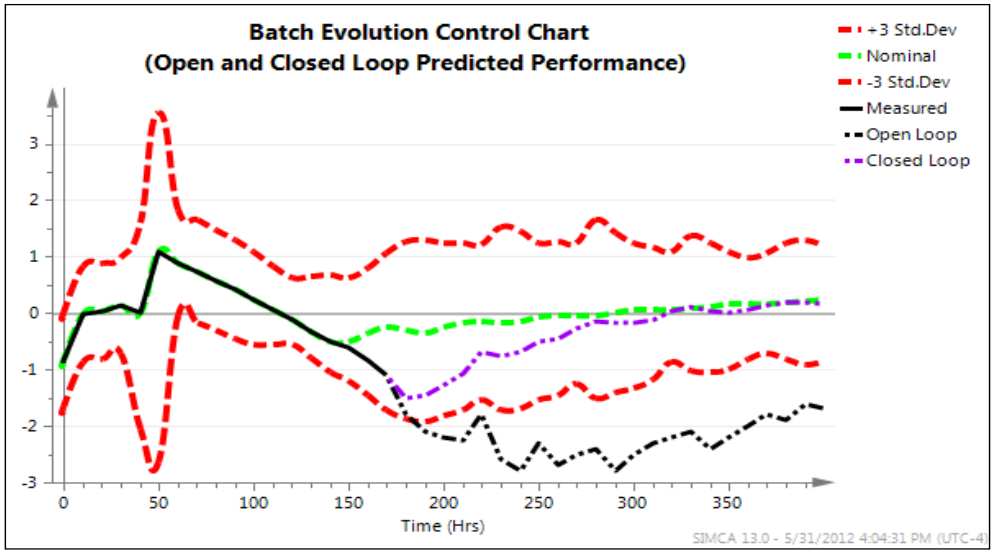


Figure 18.53: The figure depicts an example batch evolution control chart containing the information presently provided in SIMCA-online, including the measured process up to the current maturity (black), the model average (green) and +/- 3 control limits (red). With SIMCA-control these control charts may also include the desired trajectory (green dashed) as well as open loop (red dashed) and closed loop (blue dashed) predicted trajectories. Open loop refers to the predicted trajectory of the process if no process adjustments are implemented. Closed loop refers to the predicted trajectory of the process if the adjustments from SIMCA-control are implemented.

For batch level control (BLC) the goal is to find the set of X_{MV} and their settings that result in optimized final batch conditions. An example may include adjustment of nutrient levels to maximize the yield at the end of fermentation. It is envisioned that BLC is a *supervisory* type of control that is only executed at selected times for mid-course type correction. A schematic batch evolution control chart is shown in Figure 18.54.

These two methods, BEC and BLC, may be paired up in a cascade configuration to provide both *regulatory* (frequent adjustments to keep the process in-control) and *supervisory* (mid course correction) control of final qualities.

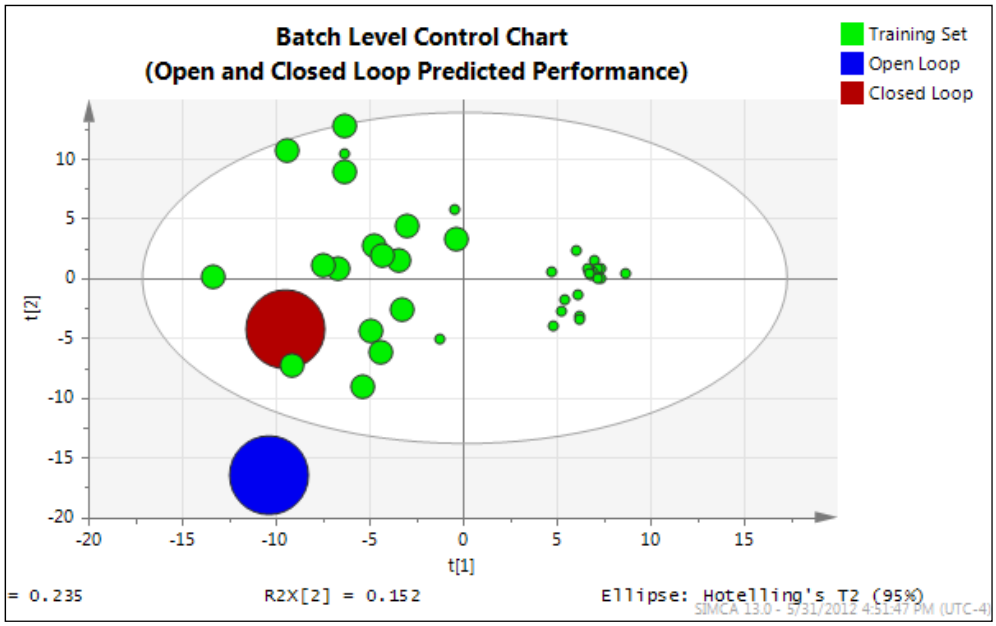


Figure 18.54: Batch level control chart showing predicted open and closed loop final batch performance.

The SIMCA-online software is used in many industries for monitoring of batch systems. It executes SIMCA offline model files. With the predictive capabilities of SIMCA-control, monitoring is enhanced by providing reliable predictions of future process performance. Model predictive control (MPC) is a technique where adjustments to a process are selected by finding the set of manipulated variables (X_{MV}) and their settings that result in an optimal process performance. Traditionally only process outputs (Y) are considered in the objective function. With SIMCA-control, multivariate terms are integrated into the objective function to assure that the adjustments made by the controller are constrained to the knowledge space [McCready, 2011]. This is especially important for regulated industries, such as pharmaceutical and biopharmaceutical, where a process must be shown to operate within regions that assure good quality product (the design space).

19 Hierarchical modeling

19.1 Objective

This chapter describes hierarchical PCA, PLS, OPLS and O2PLS. Hierarchical methods provide a flexible and versatile platform for the analysis of complex data tables, providing more transparent and interpretable results. The PROC1A and Carrageenan datasets are used for illustration. The chapter concludes with a discussion of the rationale underpinning hierarchical modeling.

19.2 Introduction

Multivariate models involving large numbers of variables are often difficult to interpret, because plots become cluttered and difficult to overview. There may then be a temptation to eliminate variables which runs the risk of losing important information and may make the modeling efforts more difficult to interpret and less robust.

Usually, a better alternative is to partition the data into blocks of logically related variables and apply hierarchical data analysis [Wold, *et al.*, 1996; Berglund, *et al.*, 1997; Rännar, *et al.*, 1998; Westerhuis, *et al.*, 1998; Janné, *et al.*, 2001; Eriksson, *et al.*, 2002; Gunnarsson, *et al.*, 2003; Stefanov, *et al.*, 2003; Eriksson, *et al.*, 2004a; Lindström, *et al.*, 2009; Hanafi, *et al.,* 2010; Hanafi, *et al.*, 2012]. In process modeling, the blocks may correspond to variables measured at different parts of the process. In spectroscopic applications, the blocks may originate from dividing the spectral window into sections corresponding to different chemical and physical information or refer to spectra from different instruments. In QSAR, the blocks might correspond to different chemical probes in 3D-QSAR or the different descriptor generator packages used to describe the molecules.

Datasets with blocks of variables like this, are conveniently analyzed by hierarchical PCA, PLS, OPLS and O2PLS. At the base level, detailed information is extracted for each different block. The scores so formed, here called ”super variables”, are combined together at the top level and analyzed in the usual way (Figure 19.1). Alternatively, the residuals from the base level models may be taken up to the top level which is known as target directed filtering of multivariate data. These and other aspects of hierarchical modeling are discussed further in section 19.6.

We also note that the hierarchical modeling approach fits nicely into the PAT framework, as promoted by the US Food and Drug Administration (FDA) and described in more detail in Chapter 18. At PAT Level 4 (Section 18.4.4), where an overview of the entire pharmaceutical production process is desired, hierarchical modeling is invaluable. This total view of the manufacturing process is called PSV, short for *process state visualization*. It embraces all unit operations and encompasses assessments of raw material, intermediates, and final products, keeping track of time trajectories of key quality and process parameters to verify that they remain within specification and are developing in the right direction.

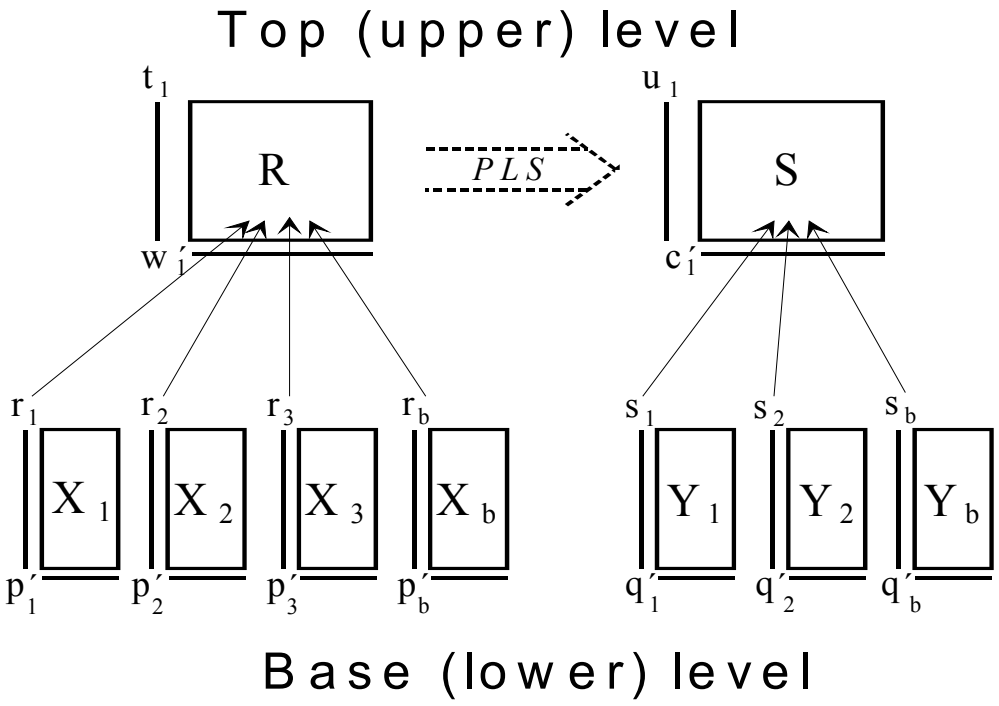


Figure 19.1: Schematic overview of hierarchical modeling. At the lower (base) level each block of data is modeled locally by either a PCA, PLS, OPLS or O2PLS model. Each block is summarized by one or more score vectors ("super variables"). Note that different numbers of "super variables" can be used for the different blocks. The computed "super variables" are then "moved" to the upper (top) level and united to form the new "X" and "Y" data matrices, depicted as "R" and "S", respectively. All conventional PCA, PLS, OPLS and O2PLS statistics and diagnostics are retained.

19.3 Process example: HI-PROC

Our first illustration of hierarchical modeling involves the PROC1A dataset which was first introduced in Section 9.5. Some additional results obtained using MSPC were presented in Section 15.7.

We briefly recall that this example deals with a polymer production plant. It is a continuous process which went out of control at around time point 80 after a fairly successful campaign to decrease the side product (response variable y6).

The manufacturing objective was to minimize the yield of side product (y6) and maximize product strength (y8). Our objective here is to investigate the use of hierarchical modeling to:

(i) overview the process

(ii) detect the process upset

19.3.1 Base level block formation

For the purpose of hierarchical modeling, we will group the variables into five blocks (Figure 19.2).

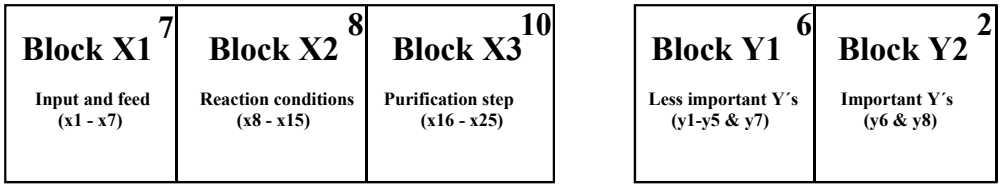


Figure 19.2: Block formation on the lower level.

19.3.2 Base level model structure

The first 69 observations (process time points) will be used as the training set to build the following four base-level models:

a) PLS of Block X1 vs Block Y2

b) PLS of Block X2 vs Block Y2

c) PLS of Block X3 vs Block Y2

d) PCA of Block Y1

19.3.3 Summarizing input and feed (1st base level model)

The PLS model of Block X1 versus Block Y2 has three components with $R^2X = 0.58$ and $R^2Y = 0.29$. Our primary objective with this model is to summarize the X-variables well. The score plot of t_2 versus t_1 (Figure 19.3) shows that the observations spread well in score space. The corresponding loading plot (Figure 19.4) indicates that y6 is modeled by the first component and y8 by the second component.

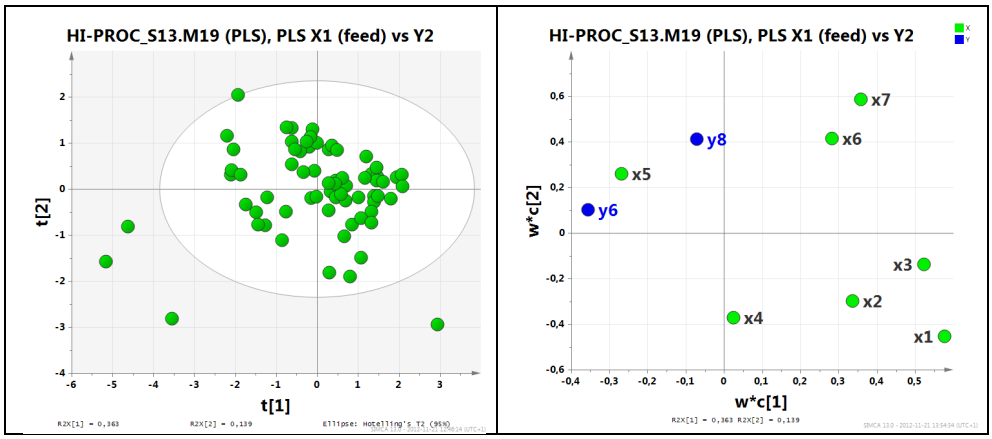


Figure 19.3: (left) PLS X-score plot (t_1/t_2). Figure 19.4: (right) Loading plot corresponding to Figure 19.3.

19.3.4 Summarizing reaction conditions (2nd base level model)

The second PLS model relates Block X2 to Block Y2. Three components were significant according to cross-validation with $R^2X = 0.73$ and $R^2Y = 0.56$. There are no outliers and the observations are evenly distributed (Figure 19.5). The PLS loading plot (Figure 19.6) suggests that y6 is related to x9, x10 & x14 and y8 to x12.

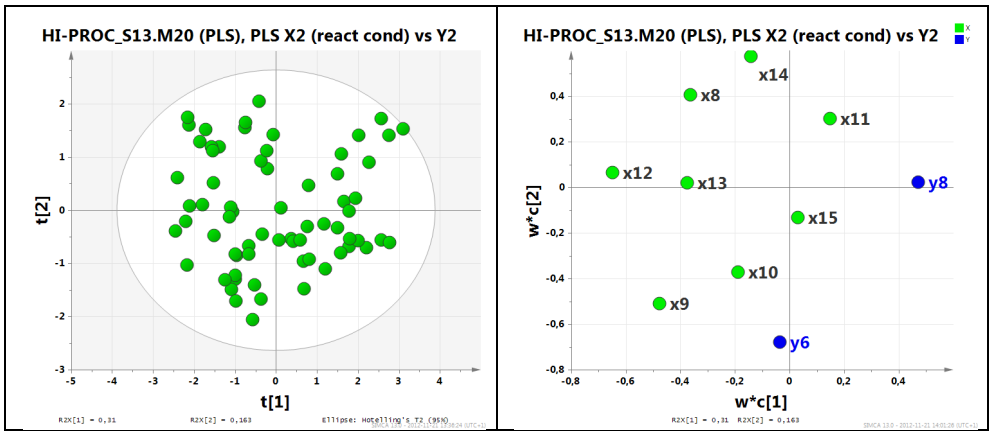


Figure 19.5: (left) PLS X-score plot (t_1/t_2). Figure 19.6: (right) Loading plot corresponding to Figure 19.5.

19.3.5 Summarizing the purification step (3rd base level model)

The PLS model relating Block X3 to Block Y2 yields three significant components according to cross-validation with $R^2X = 0.57$ and $R^2Y = 0.55$. The score plot of t_2 versus t_1 (Figure 19.7) indicates a good spread of the observations. The two responses are modeled well, y8 by the first component and y6 by the second component (Figure 19.8).

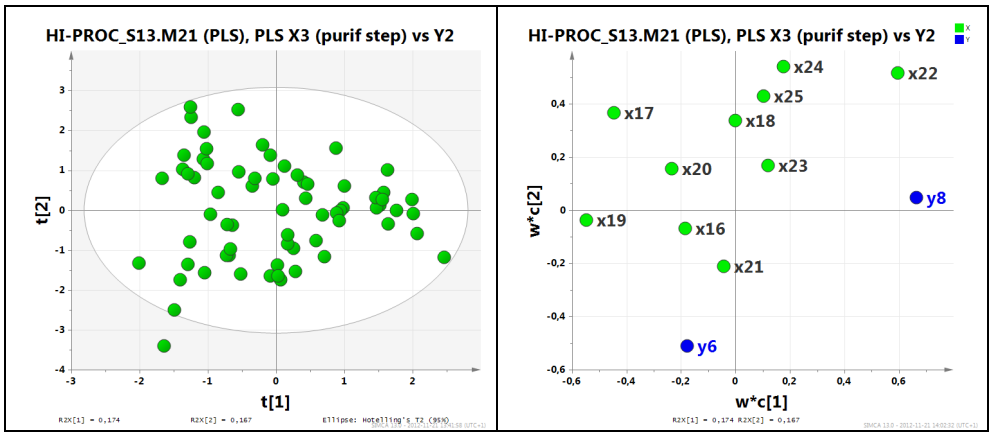


Figure 19.7: (left) PLS X-score plot (t_1/t_2). Figure 19.8: (right) Loading plot corresponding to Figure 19.7.

19.3.6 Summarizing the less important Y-variables (4th base level model)

The final base-level model consists of PCA of Block Y1, the less important response variables. There are no significant components according to cross-validation but we will calculate three components as our objective is to summarize the Y-variables well ($R^2X = 0.80$). The score plot of t_2 versus t_1 (Figure 19.9) suggests that observation 65 is an outlier in these Y-measurements. The corresponding contribution plot (Figure 19.10) indicates that the main cause for the deviating behavior is variable y7, which is much lower for observation 65 than for the average observation. Observation 65 is also strange in DModX (no plot shown) and consequently this sample ought to be deleted.

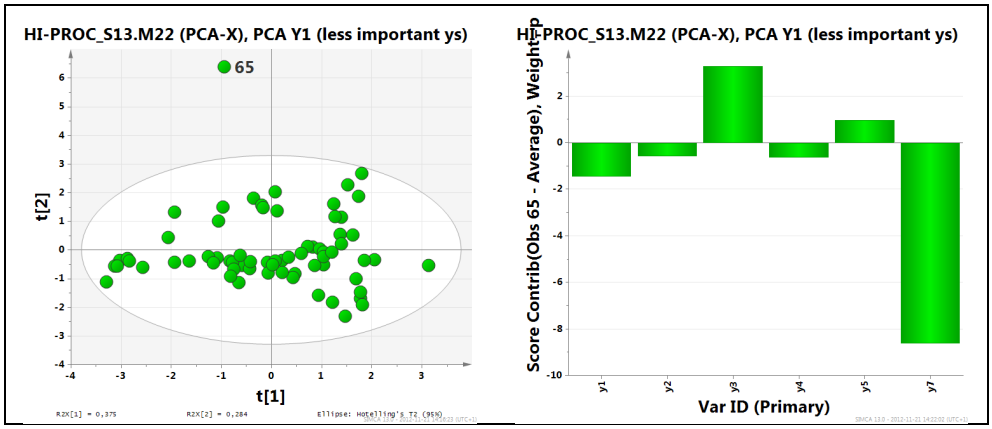


Figure 19.9: (left) PCA score plot (t_1/t_2). Observation 65 is an outlier. Figure 19.10: (right) Score contribution plot comparing observation 65 to the average observation. The contribution plot shows that variable y7 is the root cause for the problems seen for observation 65.

Observation 65 was therefore excluded and a new PCA model built for the Y1-block. Again, three components were calculated ($R^2X = 0.80$). Score and loading plots of the first two components are displayed below (Figures 19.11 and 19.12). The observations are now much more homogeneously distributed. Basically, the first component accounts for variables y2 and y4 and, to a lesser extent, y1 and y7. The second component is largely related to y3 and y5.

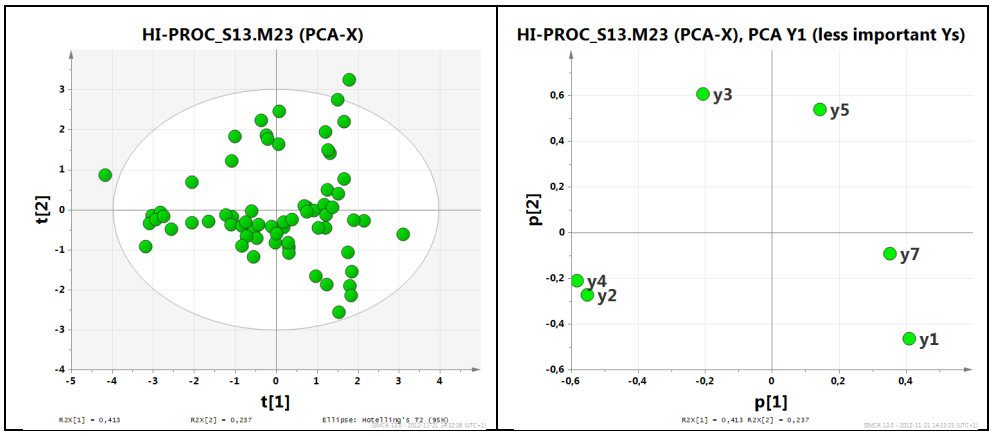
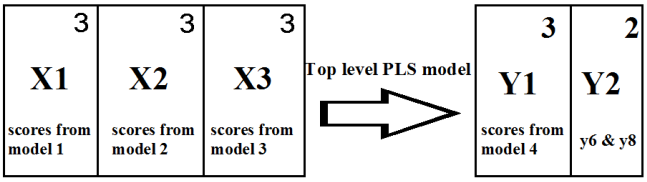


Figure 19.11: (left) PCA score plot (t_1/t_2). Figure 19.12: (right) Loading plot corresponding to Figure 19.11.

19.3.7 Top level model: PSV of the entire process

The next step is to define the X- and Y-variables for the top level PLS model. We select as the X-block the 9 score variables from the first three base-level models, and as the Y-block the two score variables from the fourth base level model plus the original variables y6 and y8. Thus, we have the data structure portrayed in Figure 19.13. All observations up to time point 69 were selected as a training set.



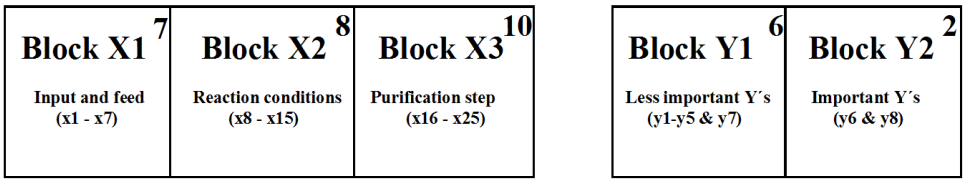


Figure 19.13: Data structure for the base and top levels in the Hi-PROC application.

The top level PLS model contained four significant components explaining 54% of the Y-variance. The first two components dominate the model (Figure 19.14). The two most important responses, y6 and y8, are modeled well (Figure 19.15) together with the first score variable (here denoted §M23.t1) of the fourth base-level model.

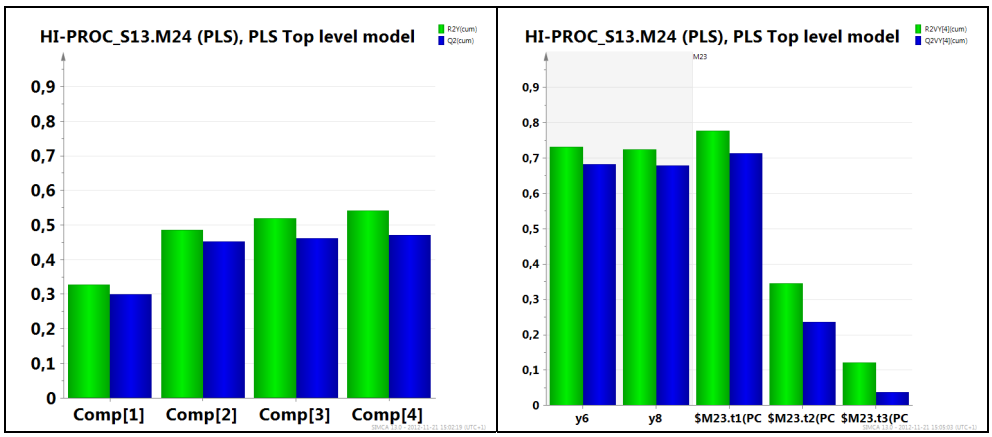


Figure 19.14: (left) Summary of fit plot of the top level PLS model. Figure 19.15: (right) Model performance for the five individual Y-variables. The notation "§M23" refers to the fourth base level model.

The score plot of t_2 versus t_1 (Figure 19.16) indicates that the process starts down to the right (high values of y6), moves to the left (lower y6), and is then manipulated to give lower values of y6 (upper left quadrant). The process then becomes unstable and lurches back to the right.

In the corresponding loading plot (Figure 19.17), y6, the side product, is on the right-hand side of the plot, inversely correlated with the first component of the feed model ($M19.t1) and the second components of both the reactor ($M20.t2) and purification models ($M21.t2). The response y6 is also negatively correlated with the first component of the model of the less important Y-variables ($M23.t1).

Meanwhile, y8 is positively correlated to the first components of both the reactor ($M20.t1) and the purification models ($M21.t1). The response y8 is also related to the second component of the model of the less important Y-variables ($M23.t2).

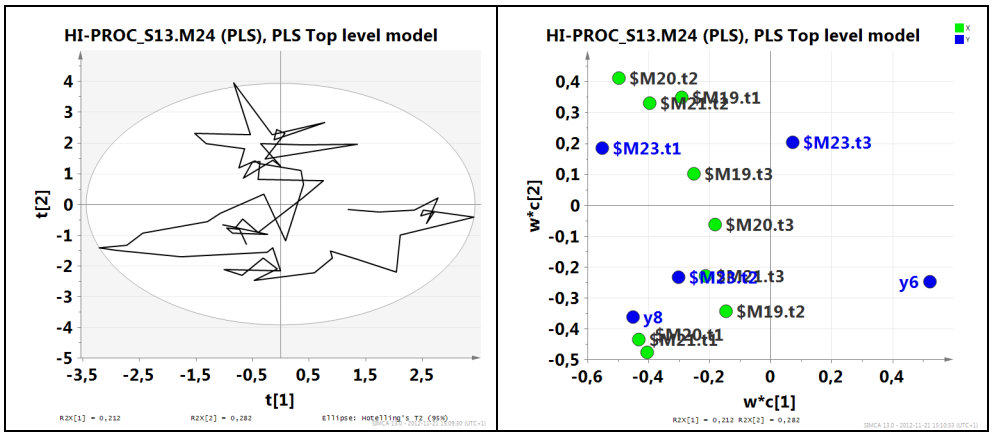


Figure 19.16: (left) Score line plot of t_1/t_2 of the top level PLS model. Figure 19.17: (right) Loading plot $w*c_1/w*c_2$ of the top level PLS model.

With the contribution tool we can investigate the patterns suggested by the loading plot above in more detail. Simply double click on any score variable (e.g. $M19.t1) and the corresponding base-level loading plot will open, revealing which of the original variables influence that particular score variable.

For example, $M20.t2 (second score vector of model for the reaction conditions) is negatively correlated with y6. Double click on this score variable and bring up the loading plot shown below (Figure 19.18). We conclude that y6 is positively correlated with x9 and x10 and negatively correlated with x8, x11 and x14.

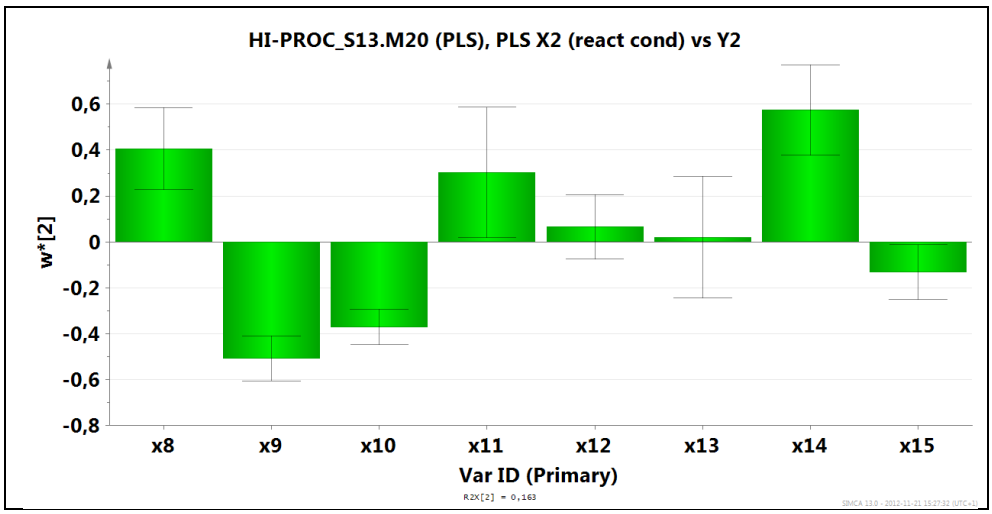


Figure 19.18: Contribution plot revealing which variables contribute to the score variable §M20.t2.

This gives us zoom-in/zoom-out functionality. In the top-level loading plot, we get a feel for the relationships between the two important y-variables and the various stages of the process (feed, reaction, purification). In the base-level loading plot, we understand which of the original variables influence these outputs.

19.3.8 Monitoring of the prediction set data

It is known that the process became unstable after time point 80, and was behaving somewhat unreliably already in the time span 70-79. In the final step, the idea is to apply the top level PLS model to the excluded observations 70 – 92 to see how well it can diagnose the process upset. Several of the prediction set time points are located outside the Hotelling´s T^2 tolerance ellipse in the predicted score plot (Figure 19.19). However, the DModXPS plot (Figure 19.20) illustrates the process upset particularly clearly. The contribution plot (Figure 19.21) for time point 91 (large DModXPS) shows the feed (model M19, components 1, 2 and 3) as being the main cause of the problem. Zooming in on the feed in all 3 components (plot only shown for t_2, Figure 19.22) points to variable 6, which is much too low. In an analogous manner, the problematic variable for observation 90 is x22 and for observation 86 x17, which are much too low (no plots shown) in both cases.

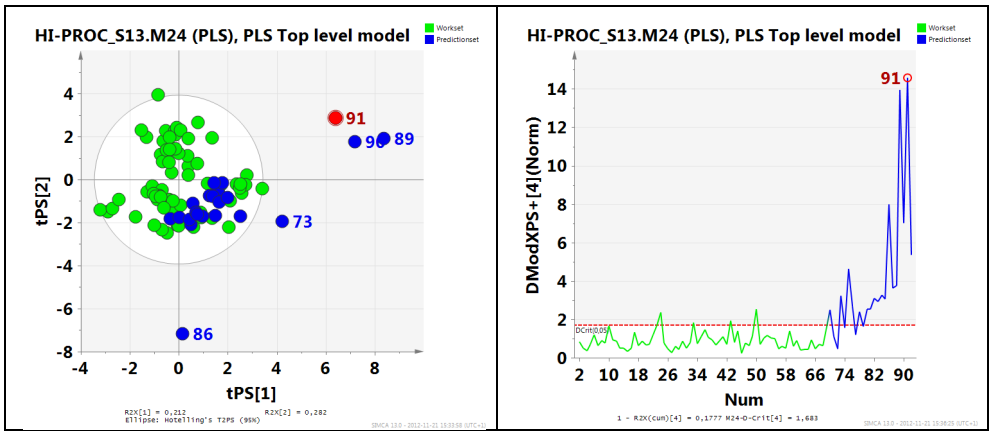


Figure 19.19: (left) Score plot of predicted scores (t_1/t_2). Figure 19.20: (right) Predicted DModX plot.

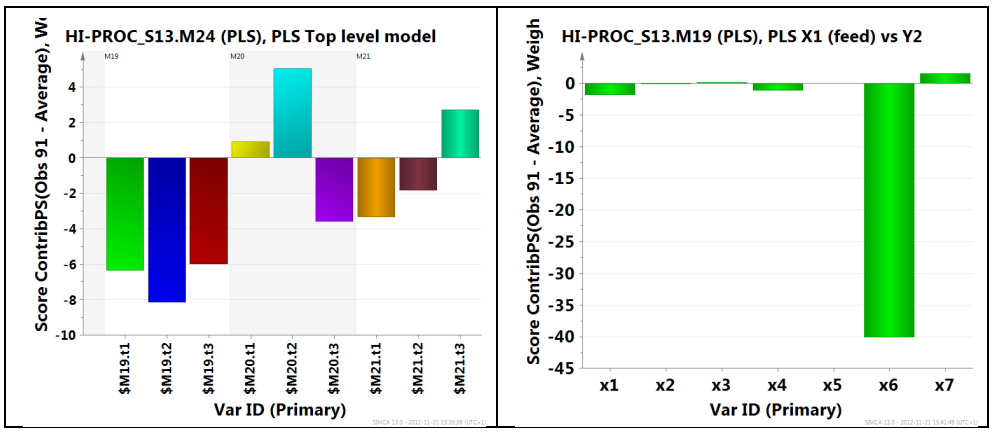


Figure 19.21: (left) Contribution plot highlighting which score variables are responsible for the deviating behavior of observation 91 in the predicted DModX plot. Figure 19.22: (right) Score contribution plot for score variable §M19.t2 showing that variable x6 is too low for observation 91.

19.3.9 Discussion of HI-PROC example

Hierarchical multivariate modeling provides the means to monitor and supervise this process. The top level score plot of t_1/t_2 elegantly shows how the process moves around in space before finally going out of control. The process upset after time point 80 was readily identified. In practice, inspection of control charts of scores, T^2 and DModX would have provided early indications of the process problems. By interrogating the various block models, deviating time points were identified and the likely cause revealed.

At the base level, a combination of PCA and PLS models were used. When the base level only comprises X-data, possibly ordered in two or more blocks, PCA is the method of choice for summarizing the data. However, when there are both X and Y data available both PCA and PLS are possible choices. The benefit of using PLS is that the model of the "X-part" is focused onto the features of immediate relevance to the Y-data. One consequence of this rotation of the "X-model" is that the top level model will tend to indicate a slightly improved Q^2, in relation to the Q^2-values obtained for the base level models.

19.4 Spectroscopy example: Carrageenan

19.4.1 Background and objective

The second example deals with a multi-block spectral dataset consisting of measurements on various species of Carrageenans. It was briefly analyzed in Section 6.5 using O2PLS. Here we will analyze it using hierarchical-O2PLS and -PCA.

Carrageenans are polysaccharides extracted from seaweed which are used as gelling and thickening agents in a wide range of industries, including food, pharmaceuticals and cosmetics. Many different types of carrageenans exist, each having different gelling and thickening properties. In commercial production, the raw material (seaweed) contains a mixture of carrageenan types and hence the final product is also a mixture of types. It is therefore very important for the industry to know the composition of their products in order to direct specific products to appropriate application areas, or, if necessary, perform chemical modification prior to release.

The primary objective of this investigation was to investigate the possibility of developing a calibration model of carrageenan composition based on spectral data acquired using near infrared (NIR), IR and Raman instruments. A secondary objective was to understand which type of spectral data provided the best basis for predicting carrageenan content in real samples. We are indebted to Jan Larsen, CPKelco A/S, Denmark, and co-workers for granting us permission to use this dataset. Original reference: M. Dyrby et al. Carbohydrate Polymers 57, 337-348, 2004 [Dyrby, *et al.*, 2004].

19.4.2 Mixture design to get representative Y-data in the training and test sets

A five-component mixture design in six levels was laid out resulting in 128 powder samples. The five Y-variables represent the relative amounts of each of the carrageenan types (Lambda, Kappa, Iota, Mu and Nu) in each mixture sample. As the five production samples used for mixing the 128 samples were not totally pure, the Y-variables are not exactly at the design levels of 0, 20, 40, 60, 80 and 100%. Furthermore, the Carrageenan types Mu and Nu are not found in natural seaweed at higher purity than about 20%. One fifth of the samples (26 samples) was set aside as a test set [Dyrby, *et al.*, 2004] leaving 102 samples in the training set (calibration set).

19.4.3 The three blocks of spectral data

For each of the 128 samples, three blocks of spectral data were acquired (see Figure 19.23):

- NIR spectra, 1100 – 2500 nm, 699 variables
- IR spectra, 550 – 4000 cm^{-1}, 662 variables
- Raman spectra, 3600 – 200 cm^{-1}, 3401 variables

As seen, the Raman block is more than five times wider than the IR block. If these blocks were modeled together, the Raman block could dominate the smaller blocks for purely numerical reasons. The SNV preprocessed spectra are plotted in Figure 19.23. The mathematical form of the SNV filter was described in Chapter 14. This filter was used to remove baseline variation.

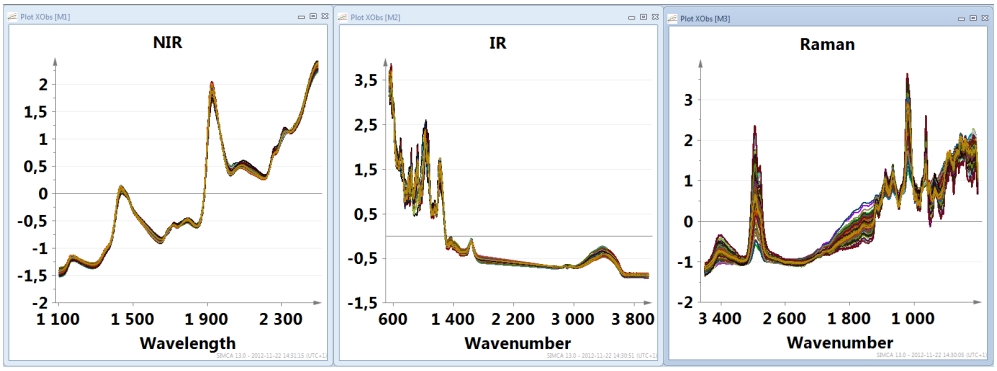


Figure 19.23: SNV preprocessed spectra of the Carrageenan example.

19.4.4 Separating Y-predictive and Y-orthogonal spectral variation

We will here demonstrate how hierarchical modeling enables us to focus the data analysis towards the structures in the spectral data that are Y-predictive and Y-orthogonal. To achieve this, we set up the models as sketched in Figure 19.24. As seen, the intention is to develop three O2PLS models at the lower level, one for each spectral block versus the same Y-block. This use of O2PLS enables a better screening of the base level score variables, i.e. which are predictive for Y and which are not. The top level modeling can then be pursued in two directions, one that further examines the Y-predictive score variables, and another which concentrates on the Y-orthogonal variation only.

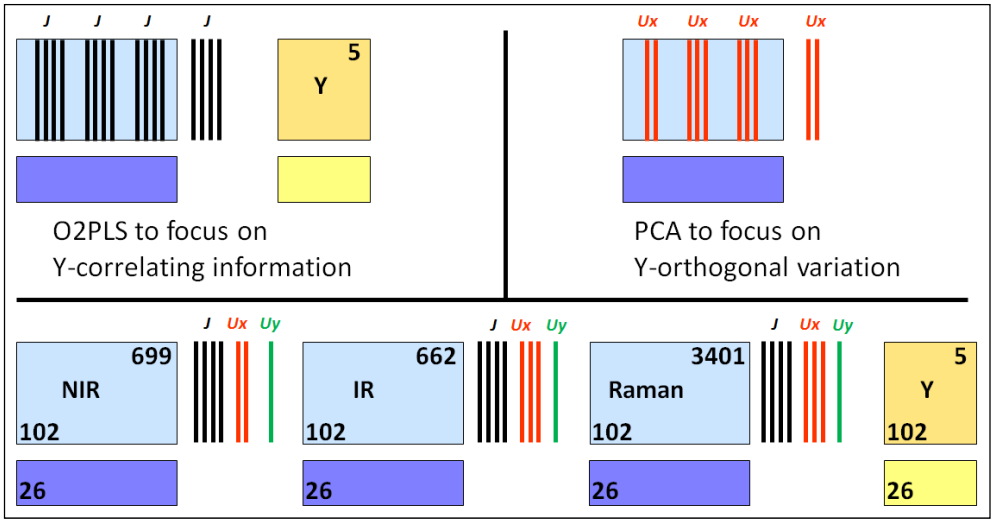


Figure 19.24: Hierarchical model structure used when separating Y-predictive and Y-orthogonal spectral variation. J means a predictive component expressing joint information. Ux means an orthogonal-in-X component expressing structure that is unique to X. Uy means an orthogonal-in-Y component expressing structure that is unique to Y.

Figure 19.24 shows that the first base level O2PLS model between the NIR block and the Y-block contains 4 predictive components expressing the joint information. Additionally, the model contains 2 orthogonal-in-X components accounting for systematic structure in the NIR data that is unrelated to the Y-block (the five concentration variables). There is also one orthogonal-in-Y component reflecting systematic variability that is unique to Y, but, given the objective of separating the spectral variability into two sources, this latter component is of lesser interest. The corresponding results for the remaining two base level O2PLS models are 4 predictive/3 orthogonal-in-X (IR block) and 4 predictive/3 orthogonal-in-X (Raman block) components.

As depicted by Figure 19.24, the predictive score variables and the orthogonal-in-X score variables were allocated to two different upper level matrices. In total, there is one dozen predictive score variables accounting for the variability in the NIR, IR and Raman data that correlates with the Y-block. The top level O2PLS model – linking the one dozen predictive score variables to the Y-block – comprised 4 components. The scores and loadings of the two first components are plotted in Figure 19.25.

The triangular structure of the underlying mixture design is easy to see (Figure 19.25). The vertices of the triangle are defined by the amounts of the Lambda, Kappa and Iota Carrageenan types. Interestingly, the loadings reveal that each of the three first predictive NIR score variables models one of these three main constituents.

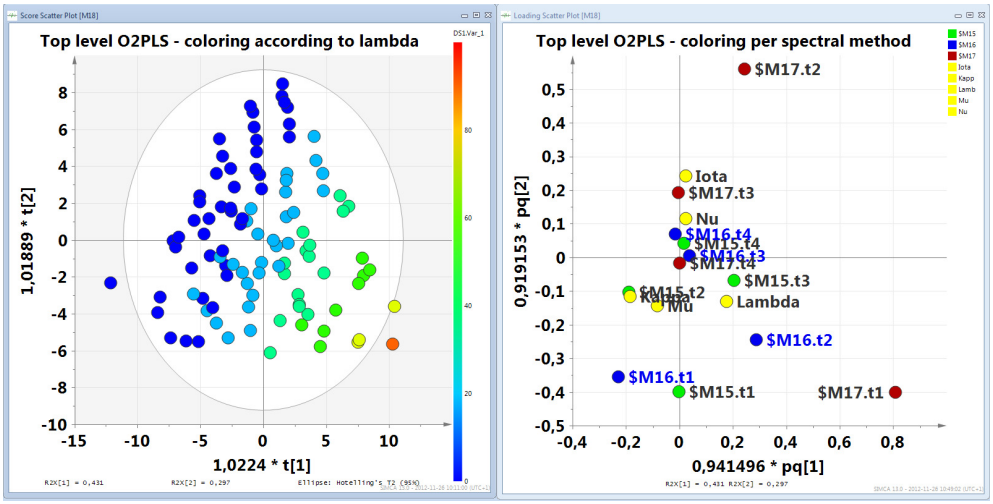


Figure 19.25: Scores (left) and loadings (right) of the first two components of the top level O2PLS model. The design structure of the underlying design is easy to see. In the loading plot $M15 refers to the NIR base level model, $M16 to the IR model, and $M17 to the Raman model.

The top level PCA model – targeting the eight orthogonal-in-X components – comprised two components. The scores of these two components are plotted in Figure 19.26 together with the first two orthogonal scores of the three base level O2PLS models. A systematic shift among the observations is seen in the top level model which is clearly due to structure in the NIR and IR data, but not Raman (Figure 19.26).

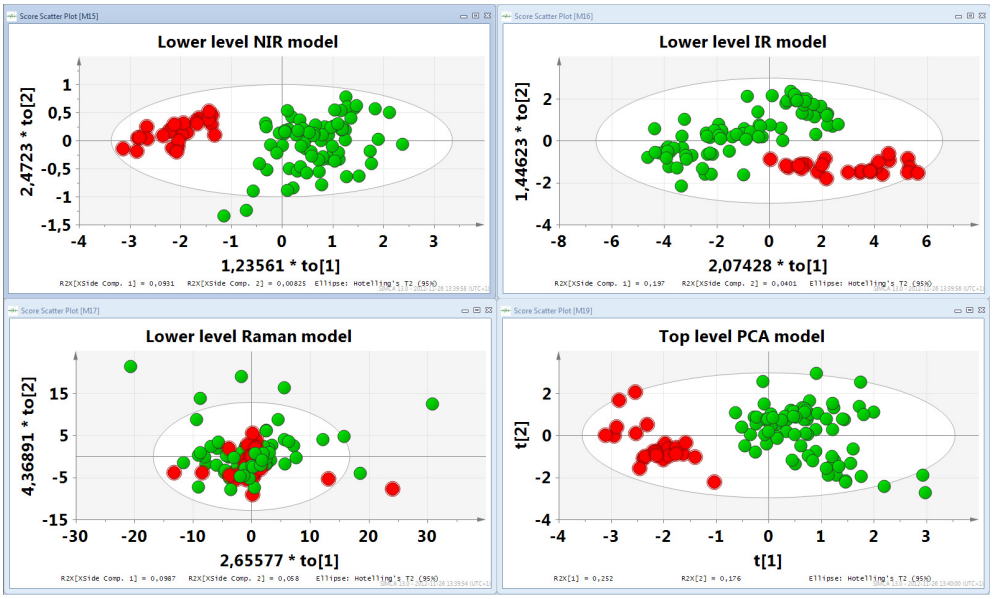


Figure 19.26: (upper left) Score plot of the two orthogonal scores of the lower level NIR model. (upper right) Score plot of the two first orthogonal scores of the lower level IR model. (lower left) Score plot of the two first orthogonal scores of the lower level Raman model. (lower right) Score plot of the two scores of the top level PCA model. The same observations (samples) are marked throughout all four plots.

To investigate further the separation along the first component of the top level PCA model, we looked at a line plot of t_1 (19.27). This line plot reveals a strong time trend in the spectral variation that is not correlated with Y. Apparently, there is a weak trend in the data during the two first days of the measurement campaign, resulting in slowly increasing score values along t_1. Subsequently, there is a jump when moving to day 3 and day 4.

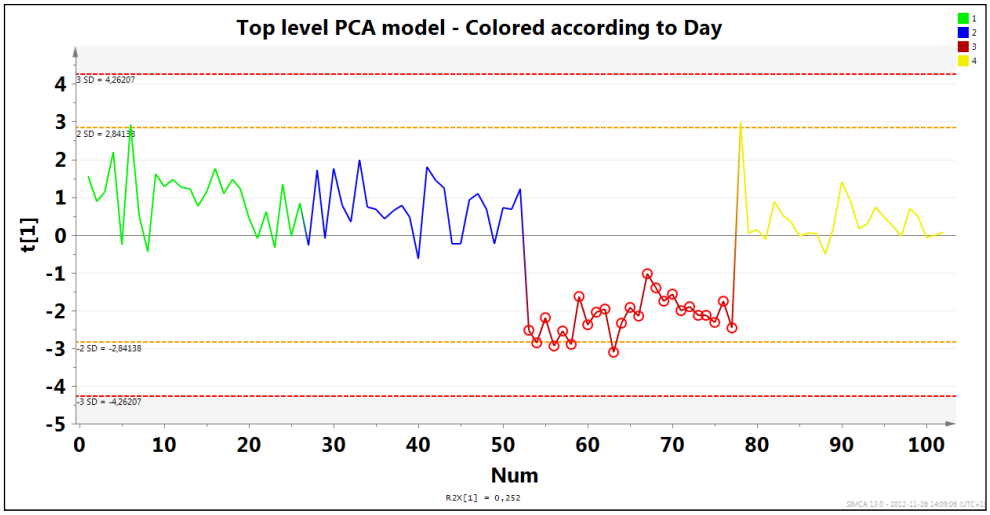


Figure 19.27: Score line plot of t_1 of the hierarchical PCA model. Systematic shifts between the different sampling days are obvious.

The group contribution plot shown in Figure 19.28 confirms that this shift is due to the NIR and IR spectra, but not the Raman spectra.

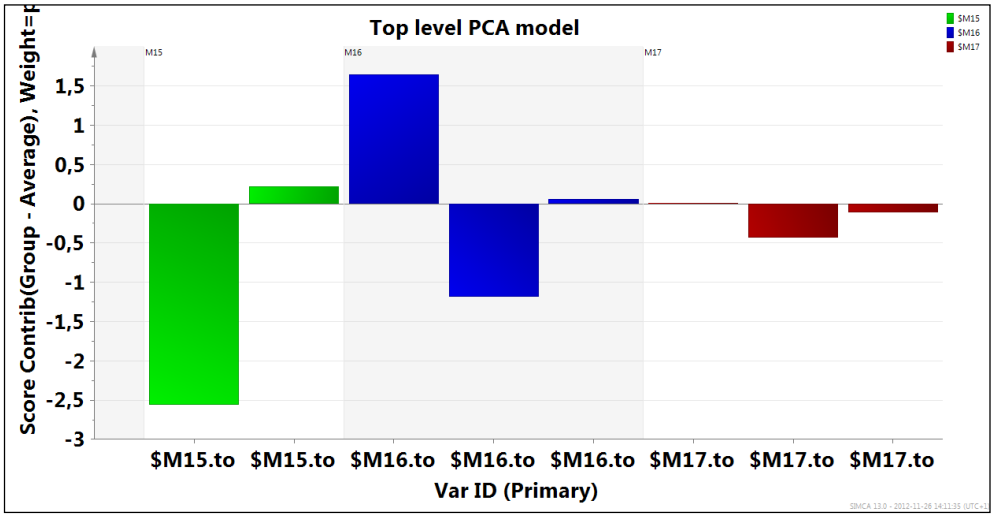


Figure 19.28: Group contribution plot between the day 3 and day 4 samples. This plot confirms that the "jump" seen in t_1 arises because of anomalies in the NIR and IR spectra.

Figure 19.29 displays how features seen in the contribution plot can be related back to the spectral domain; the line plot of the loading po_1 of the lower level NIR model suggests that the main area of Y-orthogonal spectral variation is located around 1970 nm (in the water region).

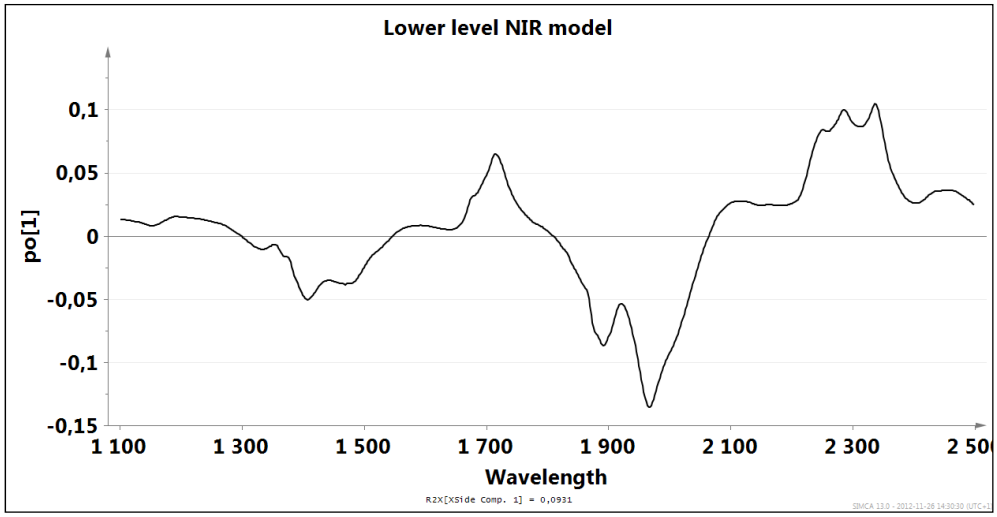


Figure 19.29: Line plot of po_1 – the loading of the first orthogonal-in-X component – of the lower level NIR model. This plot suggests the region around 1970 nm to contain a large fraction of the non-correlating variation.

19.4.5 Discussion of example

As illustrated by the Carrageenan dataset, integrating and contrasting blocks of data from different sources is greatly facilitated by the hierarchical modeling approach. The main advantages being simplicity and transparency. Other advantages are:

- blocks of very different widths are given equal opportunity to influence the top level model
- different observations can be outliers in different base level models
- the loadings plot of the top level model is less cluttered and easier to interpret but the information that the score variables represent is easily accessible.

In this example, the initial top level models point to the fact that the three X-blocks contain similar, but not identical, information. Such information redundancy is often difficult to determine in advance and so evaluating the top level model is extremely worthwhile. In other cases the different X-blocks may contain complementary information, e.g. one block of chemical data, one of texture data, and one of sensory data.

The hierarchical models thus enabled a partitioning of the spectral data into predictive and orthogonal components. A detailed analysis of the orthogonal variation revealed systematic shifts in the NIR and IR spectra between the different days of the sampling campaign. Figure 19.29 suggests this day-to-day shift to be linked to varying moisture content. NIR and IR measurements are more affected by such variations than Raman spectroscopy.

19.5 Questions for Chapter 19

1. What are the main benefits of hierarchical modeling?
2. What is meant by the zoom-in/zoom-out option?
3. How many layers can be used in a web of hierarchical models?
4. When should PCA be used? PLS? OPLS? O2PLS?
5. Which diagnostic tools help you choose the appropriate number of components?
6. Which model parameters are available for model interpretation?

19.6 Summary and discussion

19.6.1 Zoom-in/zoom-out capability

The hierarchical approach to multivariate analysis greatly enhances our ability to penetrate complex data analytical problems. The zoom-in/zoom-out capability allows us first to understand complex relationships in terms of process or system blocks, and, then, to zoom-in on individual blocks to resolve the details in terms of the underlying process or system variables. An interesting result is that the base-level models do not need that many components to capture the dominant patterns in a dataset [Eriksson, *et al.*, 2004b].

The hierarchical approach is motivated by other reasons than just enhanced model transparency and increased interpretability. First and foremost, hierarchical models are "right" from a conceptual perspective. At the base level, the analysis is based on local models which are more reliable, less prone to overfit and provide a more detailed understanding of the system or process. Global models, on the other hand, as at the top level, tend to be more superficial and can only be used to overview the most dominant patterns and trends.

19.6.2 Focus on the "right" variables at each level

One important reason why interpretation is facilitated with a hierarchical model is that the model parameters, and hence the systematic structure being captured by the combined models, are spread out across two or more model layers. The concentration of model parameters decreases at each level and, more importantly, at each level the user is directed towards the most appropriate parameters for that particular level of information in the hierarchical model pyramid.

With regard to large and ever expanding databases, we believe that the hierarchical model concept is relevant and shows great potential. Hierarchical models can be set up with more than two layers which suggests an interesting approach to scalability in which it is possible to zoom into many different model levels, thereby interpreting at different levels of detail.

19.6.3 An alternative to variable selection

The hierarchical approach is also a prudent alternative to variable deletion. In general, one should be very careful when reducing the number of variables because this can be done in so many ways that almost any result is possible. With variable selection, it is important to test the predictive power of the model on real new data rather than just repeatedly cross-validating the training set [Eriksson, *et al.*, 2003]. In practice, most X-variables contain at least some information about Y. Hence, variable selection will only result in a slight reduction in the number of variables, perhaps by 20-30 %.

Moreover, because of inherent correlations among the more important X-variables, one can continue to reduce the X-variables beyond the 20-30 % level with no apparent decrease in fit. What actually happens is that the remaining X-variables take over the role of the deleted ones, and a serious bias is introduced. This also impacts the interpretability of the models and leads to instability and reduced predictive power.

One should also remember that even seemingly unimportant variables still have a role in outlier detection. Take, for instance, a variable that is almost constant in the training set, and hence will appear as unimportant in the resulting model. If a new observation has a value for this variable that is substantially different from the training set, this indicates that the new observation is different and hence predictions of its Y-values should be treated with caution. If one mechanically deletes all variables that do not contribute to the model of Y in the training set, this automatically reduces the potential for finding outliers among new observations.

Thus, in summary, with PLS, OPLS and O2PLS and the hierarchical approach the most uninformative variables are left "untouched" at the base level. We emphasize that these variables are not removed and hence may still play a role in diagnosing outliers among

future observations. The only information that is brought up to the higher level is the variation in the relevant variables that is correlated with the Y-variables.

19.6.4 An alternative to block-scaling

One limitation of the scaling procedures discussed in Section 13.2 is that they do not take account of variables that are grouped in logical blocks, or indeed the number of variables in each block. If, for example, UV-scaling is used, a large block of variables (say, hundreds of NMR shifts) will dominate over a smaller block of variables (say, exposure conditions of animals), for purely numerical reasons. This is often not desirable.

One way of addressing this situation is to employ block-scaling (see Section 13.5). In this procedure, one may down-weight blocks of variables according to a selected scaling basis. The scaling basis chosen is generally UV-scaling, especially when the variables are of different types and have different numerical ranges. However, in multivariate calibration, no scaling or Pareto scaling may be preferred as the basis for block-scaling.

As discussed in Section 13.5, block-scaling can be done in many ways and with many basis weights and it is not always clear which method is best for a given situation. Hierarchical modeling presents a simple and flexible solution to this scaling problem. Very wide blocks of data may be summarized by a few more score vectors than the narrower blocks, which will tend to equalize the impact of the different blocks on the top-level model [see e.g. Eriksson, *et al.*, 2002]. The appropriate number of score vectors to extract from each block can be determined using cross-validation, size of eigenvalues, variance explained and other performance criteria.

19.6.5 Filtering of data

Hierarchical modeling also presents the possibility of removing confounding and/or undesirable variation from data. The first step in this procedure consists of building a base-level PLS or PLS-DA model with the undesirable feature set as Y. Then, in the second step, the residuals of the base-level model are taken to the top level and a new model is fitted to the original data with the confounding variation filtered out. This allows the user to focus on the variation of genuine interest. The ability to filter out undesired variation may be of interest in:

- Process modeling to remove e.g. the effect of machine speed or type of product
- QSAR to remove the influence of molecular weight on other descriptors
- Spectroscopy to remove temperature or pH influences on spectra
- Omics to remove species, sex, or time differences on measured data

Figures 19.30 and 19.31 show an example of eliminating strain differences seen in ^{1}H NMR-data acquired from urine samples gathered in a metabonomics project [Espina, *et al.*, 2001; Eriksson, *et al.*, 2005]. The dataset contains two strains of rat, coded as “s” and “f”. Here the focus lies on contrasting the control groups of each strain. Figure 19.30 shows the t_1/t_2 score plot from a PLS-DA model where strain membership was coded using a 0/1 representation. There is a clear separation of the two groups.

In order to focus the analysis on spectral variation beyond the strain differences, we used the X-residuals of the first model as the updated X-matrix and re-ran the PLS-DA model. As evidenced by Figure 19.31, the strain differences are removed and it is now possible to investigate other sources of variability in the NMR spectra more rigorously. As well as control animals, the same process could be used to remove strain differences from rats exposed to a drug treatment.

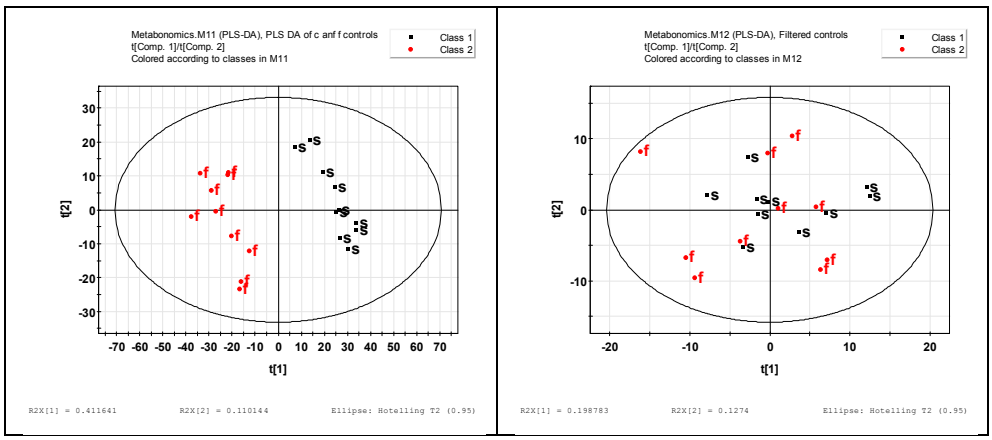


Figure 19.30: (left) Score plot t_1/t_2 of the base-level PLS-DA model. There is a distinct separation between the control animals of the "s" and "f" groups. Figure 19.31: (right) Score plot t_1/t_2 of the top-level PLS-DA model. The strain difference is removed as a result of the hierarchical filtering.

Furthermore, it should be noted that methods like OPLS and O2PLS are also available for removing systematic and undesired variation from X. With OPLS and O2PLS, the predictive component will contain the variation in X that is related to Y, while the orthogonal components will capture the remaining, Y-unrelated, X-variation. OPLS and O2PLS were described in Chapters 5 and 6.

19.6.6 Missing data and outlier handling

Hierarchical modeling also offers a flexible approach to the handling of outliers and missing data. Consider the situation sketched in Figure 19.32, in which one of the blocks (X3) has a high proportion of missing rows. If all blocks X1-X4 are treated as an entity, we may have to remove complete rows because of the abundance of missing data. Conversely, however, if the variables are blocked according to the amount of missing values, we may alleviate some of this problem. This raises the possibility of dropping observations with a high proportion of missing data in one particular base-level model whilst retaining the same observations in the other base-level models. Finally, at the top level, the problematic observations are still there but with an appropriate amount of missing data.

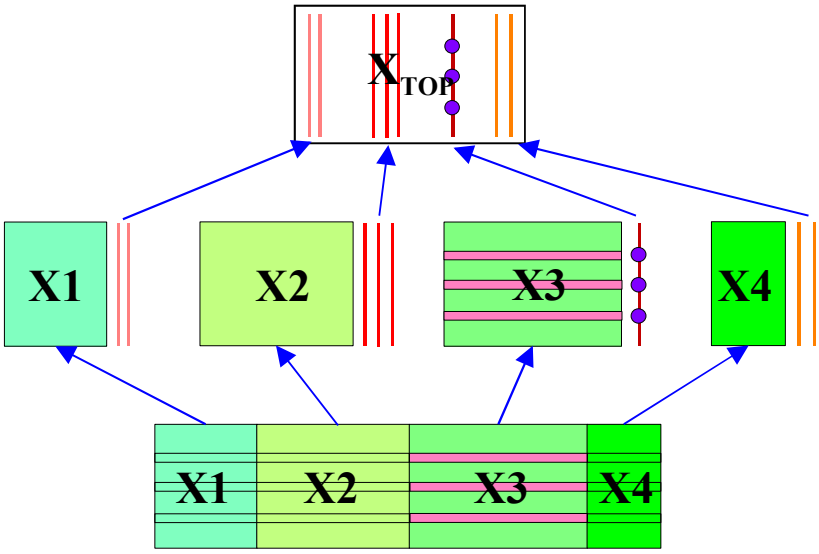


Figure 19.32: With a large proportion of missing data, we may have to remove complete rows. Blocking based on abundance of missing elements may help to alleviate this problem. In an analogous manner, the blocking approach makes it possible to remove an outlier in one particular base model whilst keeping the same observation in another base-level model.

The above approach also applies to the handling of outliers. For instance, in the HI-PROC dataset, observation 65 was diagnosed as an outlier in the final base-level model. It was omitted from that model, but not from the three remaining base-level models, and so the observation could still participate at the top-level.

19.6.7 Principal component regression, PCR

Conceptually, principal component regression, PCR, which is an alternative to PLS (see Section 4.10.5), can also be incorporated under the hierarchical modeling umbrella. The first step of PCR is a PCA of the X-matrix. In the second step, the most important score vectors are brought to the top level, where they are used as regressors in a multiple linear regression against Y. More details are found in Section 4.10.5.

19.6.8 Related multi-block techniques

Many other methods exist for the handling of blocked data where more than two matrices are considered. The BSPC method introduced and described in Chapter 16 [Wold, *et al.*, 1998b] falls into this category, as do some other (multi-way) approaches for the analysis of batch-data [Wold, *et al.*, 1987a; Nomikos and MacGregor, 1995b; Rännar, *et al.*, 1998; Westerhuis, *et al.*, 1998]. Such methods are discussed and compared in Section 16.8.4.

20 Non-linear PLS modeling

20.1 Objective

The objective of this chapter is to discuss non-linear latent variable (LV) regression analysis by means of PLS. The PLS model is linear in the relation between the two blocks X and Y. However, in many applications it is desirable to model a curved, non-linear, relationship between the blocks. The analytical calibration curve may depart from linearity, the structure-activity relationship may have a maximum or a minimum in the investigated domain, and the relationship between process variables may be non-linear. This chapter overviews a few PLS-based methods for non-linear LV regression analysis. Additionally, this chapter introduces a method for modeling and understanding non-linear relationships between X and Y, GIFI-PLS. This methodology is based on "binning" of quantitative X-variables into categorical variables. Each categorical variable is then expanded into a set of linked 1/0 dummy variables, which enable modeling of non-linearity.

20.2 Introduction

20.2.1 General considerations

PLS is a method for relating two data matrices, X and Y, to each other (see method description in Chapter 4). Usually, PLS is used for multivariate linear modeling in quantitative structure-activity relationships (QSAR), multivariate calibration, process modeling, and so on. In some cases, however, and certainly fairly often in QSAR, data structures depart from linearity. When any system or process is subjected to large changes, it appears non-linear. For instance, in QSAR, predictors and responses may sometimes extend over large ranges, and then relationships between chemical data (X) and biological data (Y) tend to be non-linear (curved). A schematic illustration of this is given in Figure 20.1. Hence, there is motivation for non-linear (PLS) modeling.

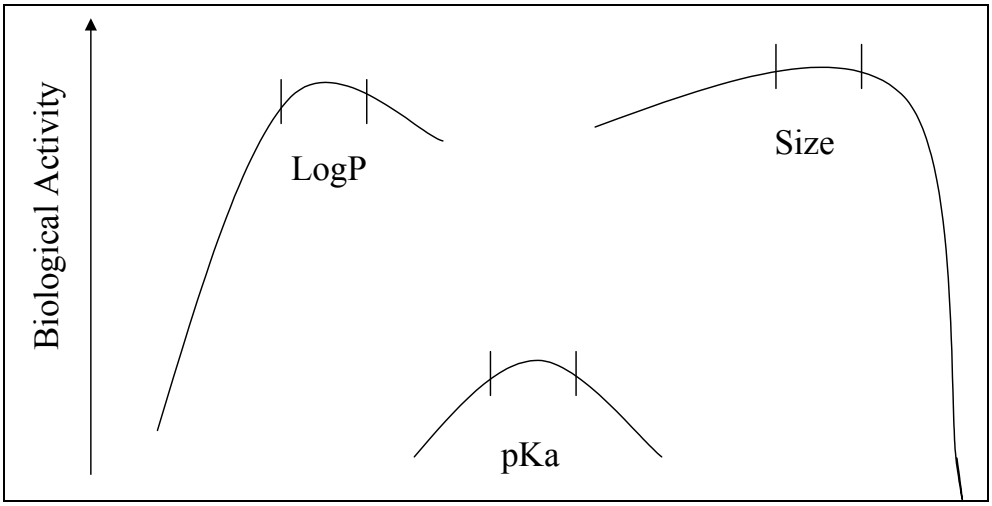


Figure 20.1: Schematic representation of the "physico-chemical window" in QSAR. In a QSAR application we are usually interested in a small region of the entire physico-chemical window. Even within such small-sized intervals often mild to moderate non-linearities exist between chemical and biological properties of compounds.

In addition to non-linear X/Y relations, the relations between the X-variables may become non-linear, as well as the relations between the Y's. Even so, the X- and Y-matrices can always be approximated by the bilinear PLS model. Hence, non-linear situations can be modeled by non-linear PLS models, where the non-linearities are expressed as non-linear relations between the X-scores (t_a) and the Y-scores (u_a). This is further discussed in Section 20.2.4.

20.2.2 Degrees of non-linearity

Relationships between X-variables and Y-variables may vary considerably in nature, as indicated by Figure 20.2. In principle, we can distinguish between several degrees of non-linearity in relationships, from mild, via moderate, to severe non-linearity where modeling attempts are problematic. The mildest non-linearities can usually be accounted for by quadratic polynomials. But the most severe non-linearities make data look discontinuous and clustered, and cannot be modeled by a single continuous model.

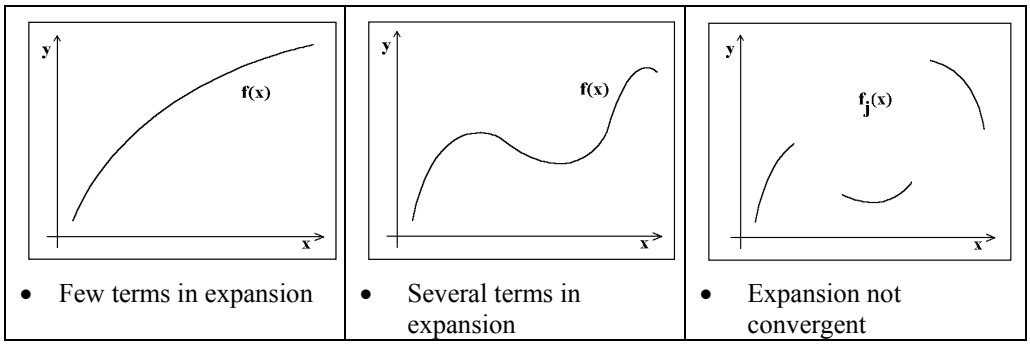


Figure 20.2: Conceptualization of different degrees of non-linearity. (left) Mild non-linearity, for which only a few higher-order terms are needed in the regression model. (middle) Moderate non-linearity requiring several higher-order terms in the expansion of the X-variables. (right) Strong non-linearity. Here, different kinds of higher order terms are usually needed for each class of observations. Hence, the expansion is not convergent. A better approach is to fit separate models for each cluster.

In QSAR, this variation in X/Y relationships is influenced by several phenomena, e.g., by the type of response modeled, the type of compound investigated, and the scope of the investigation (e.g., lead finding or lead optimization). For simple cases with unspecific and unselective endpoints, e.g., acute toxicity to fish, linear QSAR models are usually obtainable that may extend over several classes of compounds [Hermens 1989; Könemann, 1981]. For more selective and specific responses, however, response/predictor relationships may deviate from linearity, and mild to moderate non-linearity is seen in many cases [Kubinyi, 1990; Schaper 1991; Seydel, *et al.*, 1991; Kim, 1993; Frank, 1995]. Situations involving mild to moderate non-linearity may often be handled with some type of non-linear regression modeling.

Non-linearity may also occur in process data. In controlled processes, where key variables are usually allowed to vary only over small intervals (Δ small), i.e., the set point $\pm \Delta$, the typical result is a problem of linear or near-linear nature. However, sometimes in partially controlled processes, Δ may be large for some key variables. A typical example may be a waste water treatment facility in a process where variables like feed flow, pH, and Cl_{total} may vary widely, and hence influence important environmental parameters like COD (chemical oxygen demand) non-linearly. Relationships between process variables, including input/output relationships, then tend to become non-linear.

Moreover, in some cases extreme non-linearity is encountered, which cannot be adequately and parsimoniously represented by a single non-linear regression model. Such extreme circumstances may arise in QSAR when addressing selective and specific biological or environmental responses, and strongly clustered compounds, i.e., cases where chemical property spaces contain several dense regions, clusters, of compounds with empty space in between. Here, relationships are so non-linear that serial expansions diverge, and the only recourse is to divide the data into more homogeneous groups (clusters) and model each

group separately. Note that two approaches to clustering of large datasets are discussed in Chapters 21 and 22.

20.2.3 Approaches to diagnosing and handling non-linearity

As was discussed in Chapter 4, Section 7, PLS contains a number of useful diagnostic tools. In particular, we highlighted the diagnostic power of the inner relation, i.e., the t/u score plots. A t/u score plot detects at least strong non-linearity, if present, as inferred from a curved distribution of the observations (see Figures 20.3 and 20.4).

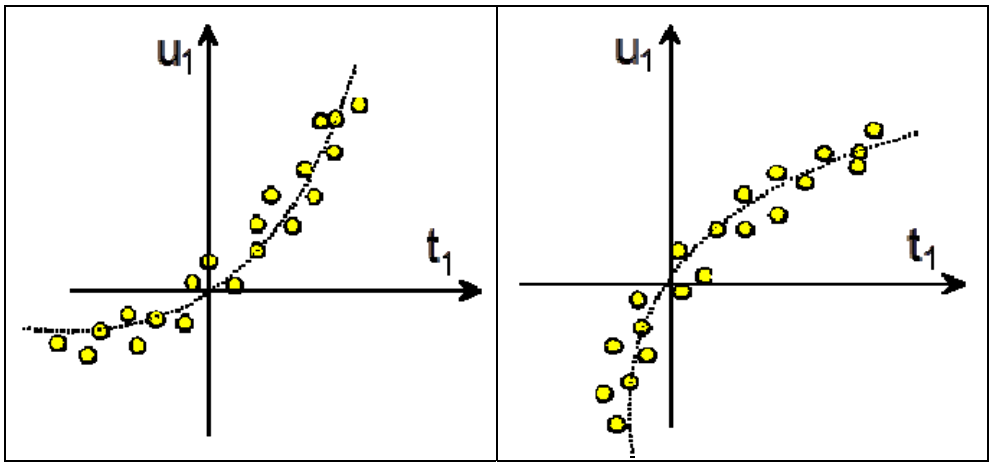


Figure 20.3: (left) Non-linearity detected by the first PLS component. By log-transforming the Y-data, the "concave" inner relation in the PLS model becomes more nearly linear. Figure 20.4: (right) Other ("convex") non-linearity seen in the first component of a PLS model. Here, log-transformation of the X-data may linearize the regression problem.

The curved appearance of an inner relation is an indirect indication of non-linearity. Many variables may contribute more or less strongly to a PLS component, and regression coefficients or PLS weights/loadings may have to be painstakingly evaluated before the non-linearity is fully understood. In contrast, the GIFI-PLS method is much more direct in the exploration of non-linearity. This methodology allows rapid assessment of whether the non-linearity in the X/Y-relation is confined to just a few variables, or whether it is a more general phenomenon expressed by most of the variables. GIFI-PLS is further discussed in Section 20.3.

One approach that may sometimes eliminate non-linearity is transformation of X and/or Y. In Chapter 4, the logarithmic transformation was shown to be useful for linearizing non-linear QSAR and process problems. Figures 20.3 and 20.4 indicate what the PLS inner relation may look like when encountering non-linearity. Depending on whether the curvature is "concave" or "convex", applying the logarithmic transformation to either Y or X will often produce a less non-linear relationship.

Non-linearity may also be handled by non-linear regression methods. The simplest non-linear regression approach is to just expand the X-matrix with the squares or cubes of its columns, and then use this expanded matrix to model Y by PLS. Classical non-linear PLS models are other alternatives, where the non-linearities are expressed as non-linear relations between the X-scores (t_a) and the Y-scores (u_a). GIFI-PLS is yet another alternative. These methods and others are further addressed in the next section.

20.2.4 Non-linear PLS and other non-linear modeling tools

Numerous approaches have been proposed for non-linear regression modeling, including methods like alternating conditional expectations (ACE), multivariate adaptive regression splines (MARS), neural networks (NN), genetic algorithms (GA), support vector machines (SVM), non-linear PLS (NPLS), spline PLS (SPLS), locally-weighted PLS (LWPLS) and hybrids between PLS and neural networks [Wold, *et al.*, 1989b; Wold 1992; Qin and

McAvoy, 1992; Sekulic, *et al.*, 1993; Andersson, *et al.*, 1996; Blanco, *et al.*, 2001; Leardi, 2001; Norinder, 2003; Czekaj, *et al.*, 2005; Searson, *et al.*, 2007; Kohonen, *et al.*, 2009; Li, *et al.*, 2009; Michielan and Moro, 2010; Kim, *et al.*, 2011; Nakagawa, *et al.*, 2012; Töndel, *et al.*, 2013]. Neural Networks, MARS, ACE, GA and SVM are examples of non-linear regression methods that deal with mild to moderate non-linearities. These methods work well for long and lean data structures where N is large (many observations) and K is small (few X-variables), i.e., when $N > 10\ K$, and $K < 20$. However, they are less suitable for short and wide datasets. Several of these methods are regression-type approaches explicitly or implicitly using matrix inversions. Concerning the NN's, these are also difficult to interpret (non-unique solutions).

There are several ways to accomplish non-linear PLS modeling. The simplest polynomially non-linear approach is to just expand the X-matrix with higher order terms, and then use this expanded matrix to model Y by PLS. Berglund and Wold [Berglund and Wold, 1997a; Berglund and Wold, 1999] showed that in mildly non-linear cases, it is sufficient to include only quadratic terms of the X-variables. This was called implicit non-linear latent variable regression, INLR. INLR has been shown to work well in calibration and QSAR problems, where the degree of non-linearity is mild. INLR was discussed in Chapter 12.

Also, note that we do not here deal with the much simpler problem when an experimental design supports a quadratic regression model [Box, *et al.*, 1978]. Design of experiments (DOE) is useful in the planning of representative experiments, and such experiments can be tailored to fit interaction, quadratic, or even cubic regression models. PLS can then be used accordingly, and this is often called PLS-RSM (response surface modeling). PLS-RSM was discussed in Chapter 18.

Other approaches to non-linear PLS are based on using a non-linear inner model in PLS (i.e., $t_a \rightarrow f(t_a) \rightarrow u_a$). Such non-linear phenomena can be modeled as polynomial non-linearities (quadratic, cubic, etc.), spline functions, or other non-linear forms, e.g., bi-exponential. Figure 20.5 displays an illustration of spline-PLS. This kind of polynomially expanded inner relation PLS model applies to mild and moderate non-linearities. These PLS approaches experience the same problems as do other non-linear models (e.g., neural networks), i.e., that they work well when N is large (many observations), but are otherwise much too flexible and adaptable. Outliers, jumps, and other anomalies are all easily accepted by the model.

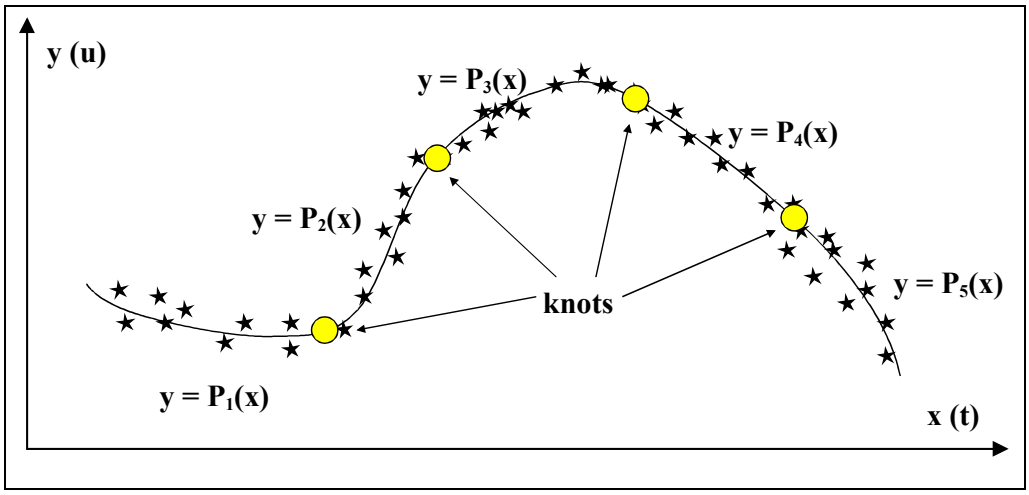


Figure 20.5: Schematic of Spline-PLS. Spline functions are piecewise polynomials joining at knots with continuity constraints. The spline function is used to model a non-linear relationship between the X- and Y-scores. This approach works well when N is large (many observations) but is otherwise much too flexible, i.e., it adapts to anything. Outliers, jumps, …, are all "accepted" by the model. The same problem applies to other non-linear models, e.g. neural networks.

In summary, the mildest non-linearities are those that can be modeled by quadratic polynomials. Here, INLR, quadratic PLS, and all other non-linear approaches work well.

Moderate non-linearities which require higher-order polynomials can be modeled by NN-PLS and spline-PLS. The most severe non-linearities, however, make the data look discontinuous and clustered, and cannot be modeled by a single continuous model. How to handle strongly clustered datasets is further discussed in Chapters 21 and 22, which present tools for top-down and bottom-up clustering situations.

20.3 Binning and expansion of X-variables (GIFI-PLS)

20.3.1 Introduction to GIFI-PLS

We will now describe a simple and useful approach to non-linear regression modeling, GIFI-PLS. The idea of GIFI-ing is simple, and hence powerful. X-variables (or X-scores) that are suspected to have non-linear relations to Y are divided into bins, and each bin is thereafter represented as one 1/0 dummy variable. The GIFI-approach has been used together with PLS on the modeling of toxicity data and process data [Berglund, *et al.*, 2001; Eriksson, *et al.*, 2000].

It is noted that the GIFI-system of multivariate data analysis has been applied in various fields in the physical, social and life sciences [Michailidis and de Leeuw, 1998]. As discussed by Michailidis and de Leeuw, any continuous numerical variable can be thought of as a categorical variable with a very large number of categories. In other words, any continuous variable, including scores, may be handled by the GIFI-system.

20.3.2 Main steps of GIFI-PLS

We will now illustrate the main steps for converting a continuous variable into a "GIFI-ed" variable.

1. *Range*. The first step corresponds to determining the range of the (X-) variable (Figure 20.6a).

2. *Number of bins* (G). Then an appropriate number of bins (G) is selected for each variable.

3. *Binning*. The data points are then binned according to the different intervals (Figure 20.6b).

Expansion to 1/0 dummy representation. Finally the binned data are re-coded as a set of "qualitative" (categorical) variables (Figure 20.6c).

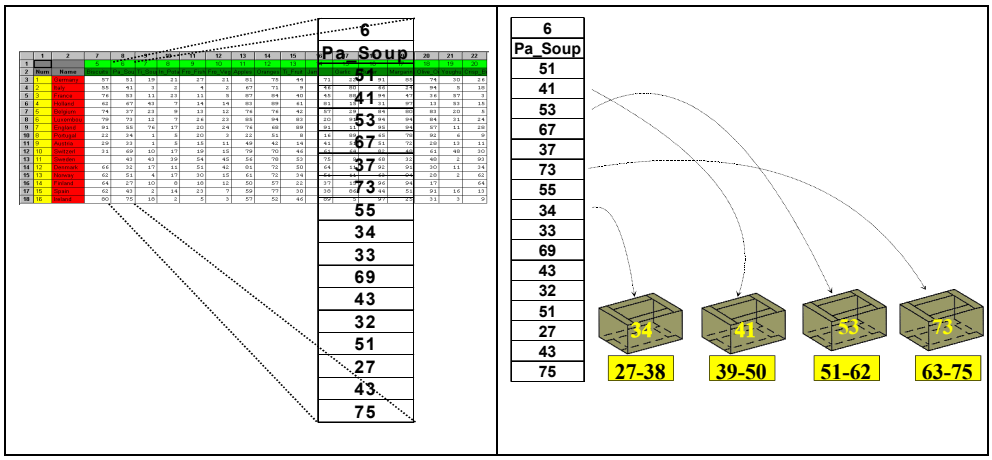


Figure 20.6: Illustration of the steps involved in the "GIFI-ing" procedure. (a, left) Step 1: The range of the X-variable is computed. (b, right) Step 2: An appropriate number of bins (G) is selected for each variable, and the data points are then binned according to the different intervals and re-coded as a "qualitative" (categorical) variable./continued on next page/

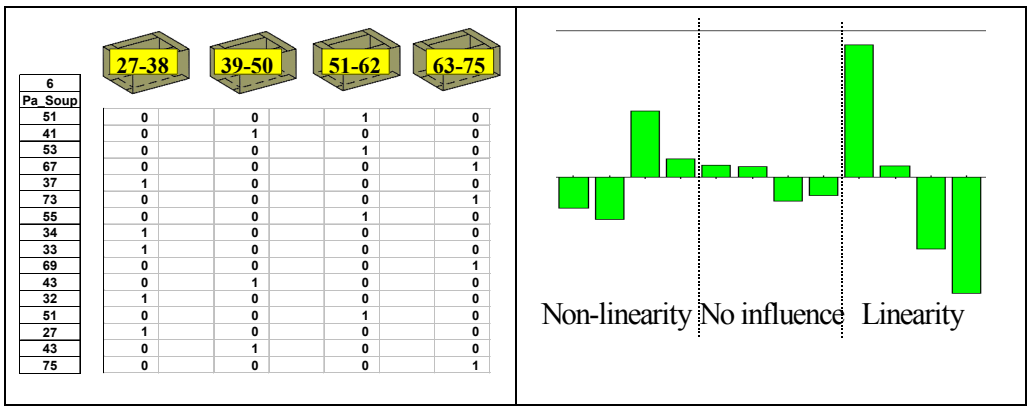


Figure 20.6: Illustration of the steps involved in the "GIFI-ing" procedure. (c, left) Step 3: All the "qualitative" (binned) variables are expanded into sets of (1/0) dummy variables, which are then used as X-variables in the PLS model. (d, right) Step 4: In the interpretation step different regression coefficient profiles will indicate various relationships between the X- and the Y-variables. Different cases of linearity, non-linearity, and an uninfluential X-variable, are given.

Subsequently, all the "qualitative" (binned) variables are expanded into sets of (1/0) dummy variables, which are then used as X-variables in the PLS model (Figure 20.6c). This binning and expansion of quantitative variables enable non-linearities to be captured. The ensuing model interpretation in terms of patterns among binned sets of regression coefficients will reveal non-linearities. Figure 20.6d illustrates different regression coefficient patterns – for a four-binned variable – indicating situations of linearity, non-linearity, and a non-influential X-variable.

20.3.3 Properties of GIFI-PLS

To understand the effect of binning we will work out a simple example. Consider a variable pH varying between 1.5 and 5.2. This may be divided into five bins with the borders 1.50, 2.24, 2.98, 3.72, 4.46, and 5.20. A pH value of 1.71 falls in bin 1, a value of 4.92 falls in bin 5, etc. Instead of using pH in the data analysis, we now use a set of five 1/0 variables, one for each bin.

The binning procedure decouples the linear order between the original quantitative values, and allows any pattern in the relation between pH and Y to be modeled. The pattern in the resulting set of coefficients – one coefficient for each pH-bin – will indicate if the relationship between pH and Y is linear, or approximately quadratic or cubic, etc. Thus, the effect of GIFI-ing is a loss of resolution and precision, but a gain in model flexibility. In modifications of GIFI-ing discussed below, representations other than 1/0 dummies are used. By using the product of the original data-values and the 1/0 dummies, one achieves a more continuous modeling than with the simple 1/0 representation.

20.4 QSAR Example: ELASTASE (Peptide QSAR)

20.4.1 The data

This example deals with peptide QSAR and a dataset comprising 89 elastase substrates. The general peptide formula is Suc-Pos1-Pos2-Ala-pNa, with Pos1 and Pos2 being the two varied positions, and Suc and pNa representing succinyl- and p-nitroanilide moieties, respectively. Two response variables are available, log k_{cat}, and log k_{cat}/K_m. These are kinetic parameters relating to substrate binding to enzyme (K_m) and to formation of acylated enzyme and release of p-nitroaniline (k_{cat}) [Nomizu, *et al.*, 1993]. This dataset is known to contain non-linear relations, see e.g. [Kimura, *et al.*, 1996]. Moreover, since the ELASTASE dataset contains some non-coded amino acids an extended set of amino acid scales that encompasses 87 amino acids was utilized [Sandberg, *et al.*, 1998]. This set contains five scales, denoted zz_1-zz_5.

20.4.2 Reference PLS model

The ELASTASE (linear) reference PLS model comprised 2 components (A = 2), displaying $R^2X = 0.29$, $R^2Y = 0.59$ and $Q^2Y = 0.48$. The Q^2Y-values for the individual responses were 0.53 (log k_{cat}) and 0.43 (log k_{cat}/K_m). Figure 20.7 provides the PLS score plot of the first dimension. We can see that the "correlation band" is rather wide. The PLS weights (Figure 20.8) show that log k_{cat} is mainly influenced by zz_1 (hydrophobicity) and zz_3 (polarity) of position 1. There is also a strong contribution from zz_5 of position 2. The novel zz_5 scale is also influential -- in both positions -- with regard to the other response, log k_{cat}/K_m. For a more thorough interpretation, reference is given to [Sandberg, *et al.*, 1998].

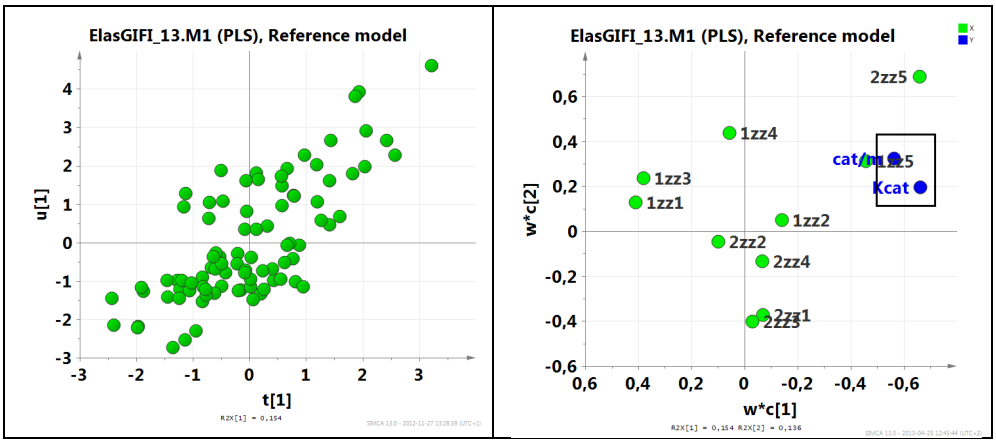


Figure 20.7: (left) PLS scores of the first dimension of the reference model. This component models 50% of the variation in biological activity. Figure 20.8: (right) PLS weights of the reference model. The variables are: zz_1-zz_5 of position 1 (1zz1-1zz5) and zz_1-zz_5 of position 2 (2zz1-2zz5). The two responses are boxed.

20.4.3 GIFI-PLS model

The 10 X-variables of the ELASTASE dataset were binned using five intervals, i.e., a total of 50 1/0 dummy variables were formed. Interestingly, the GIFI-PLS model was much better than the reference model, as $R^2X = 0.18$, $R^2Y = 0.84$, and $Q^2Y = 0.76$ (A = 2). The substantially improved separate Q^2Y-values of 0.77 (log k_{cat}) and 0.75 (log k_{cat}/K_m) confirm that the binning procedure has enabled non-linearity to be picked up for both responses.

Figure 20.9 indicates that the binning procedure has given a more regular structure to the inner relation. However, the model interpretation (Figure 20.10) is a little more demanding since we now have 50 1/0 dummy descriptors to consider. Avoiding a deep interpretation, we note that the binned variables of zz_1, zz_3, and zz_5 dominate in both amino acid positions.

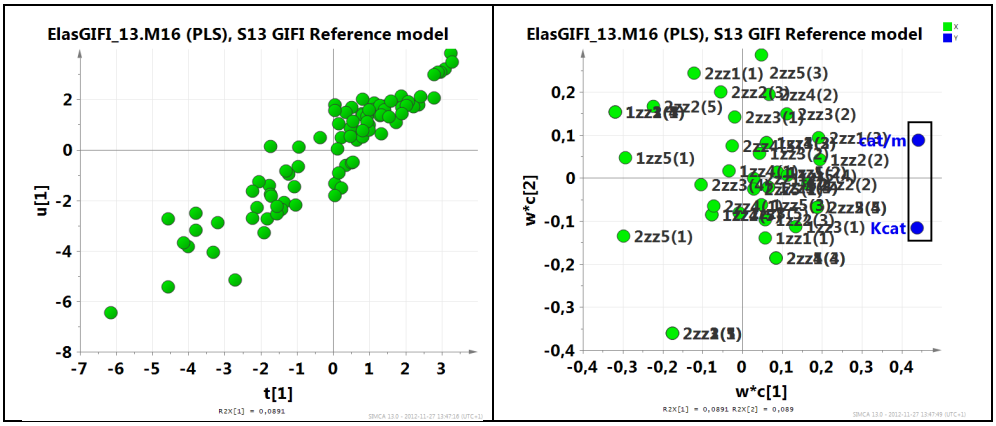


Figure 20.9: (left) PLS scores of the first dimension of the GIFI-PLS model. This component models 69% of the variation in biological activity. Figure 20.10: (right) PLS weights of the GIFI-model. In the name of a variable, the first four characters indicate the precursor descriptor, and the value in parentheses indicates the bin number. The two responses are boxed.

20.4.4 Full PLS model

In this case, the pooling of the two sets of descriptors (original and binned zz-scales) added little to the modeling. This is because $R^2X = 0.30$, $R^2Y = 0.85$, and $Q^2Y = 0.76$ (A = 3), i.e., values that are strikingly similar to the GIFI-model, but which demand one extra PLS component. Hence, the set of GIFI-descriptors is preferred to the combined pool of predictors.

20.4.5 External predictivity assessment

In the publication of Sandberg *et al.* [Sandberg, *et al.*, 1998], a D-optimal subset of 32 peptides (with numbers 1-32) was identified as the training set. We decided to use this training set. Consequently, the remaining 57 peptides (with numbers 33-89) were employed as an external prediction set. We also decided to preserve the binning procedure established above.

Two new QSARs were then computed, one using the original and one using the binned zz-scales. When using the zz-scale model, an overall external RMSEP (root mean square error of prediction) of 0.86 was achieved. The external RMSEP's of the two responses were 0.38 (log k_{cat}) and 0.48 (log k_{cat}/K_m). The corresponding RMSEP values for the GIFI-model were: 0.59 (overall), 0.29 (log k_{cat}) and 0.30 (log k_{cat}/K_m). In summary, the GIFI-approach enabled a lowering of the prediction error by approximately 31% (computed as 1 – ((0.29/0.38 + 0.30/0.48)/2).

20.4.6 Summary

The ELASTASE dataset was treated according to the principles of peptide QSAR (discussed in the supplementary book material available on www.umetrics.com). However, because this dataset contained some non-coded amino acids, an extended set of five amino acid scales, the zz-scales, was employed. It was shown that the GIFI-approach produced better external predictions than linear PLS-modeling. The GIFI-PLS model enabled the prediction error (RMSEP) to be lowered from 0.86 (zz-scale model) to 0.59 (GIFI-model), i.e., a decrease by approximately 31%.

20.5 Process Example: SimCODM

20.5.1 Introduction to SimCODM

At the Aylesford Newsprint papermill in Kent, UK, premium quality newsprint is produced from 100% recycled fiber, i.e. reclaimed newspapers and magazines. The collected paper consists of approximately 70 % used newspapers and 30 % used magazines. After collection, the paper is delivered to the recycled fiber store at the mill. A controlled proportion of the recycled paper is then loaded onto two conveyor belts feeding the deinking process.

The wastewater discharged from the deinking process has been identified as the main source of COD (chemical oxygen demand) in the mill effluent. The COD is a routinely measured parameter that estimates the oxygen requirement of oxidizable organic matter present in the effluent. The COD in the effluent is reduced by 85 % in the effluent treatment process prior to discharge to the river Medway. Stricter environmental regulations and costs associated with the reduction of COD are, however, encouraging the control and reduction of COD at source.

One of the company's environmental objectives was to investigate sources of COD generation and to identify control and reduction strategies for the future. The multivariate approach was adopted as a way of achieving this [Eriksson, *et al.*, 2001]. The objective of the current study was two-fold, i.e., (i) to gain insight into the dominating properties of the

process, and (ii) to reveal the main factors affecting the level of COD in the final effluent. Below, we will review the main steps of the work undertaken.

In order to get a good overview of the process properties, data from more than one year of production were used for the multivariate study. The dataset contains 384 observations, one response variable (COD load, y1) and 54 process variables (x2 – x55). Variables x2 – x11 relate mainly to brightness and residual ink measurements in the process; variables x12 – x20 relate to conditions at the rotating drums and the pre-flotation; variables x21 – x25 represent addition of specific process chemicals (e.g., silica, caustic and peroxide) to the dispergers; variables x26 – x37 reflect process conditions (e.g., sludge flow, effluent flow, polymer addition) in the flotation and thickening stages; variables x38 – x39 describe pH in incoming and final effluent; variables x40 – x51 correspond to process temperatures registered at various places in the deinking plant; variables x52 – x54 reflect measured effluent parameters, and finally variable x55 signifies the pulp production to the DIP-towers.

Because (a) the composition of the recycled paper and (b) the feed to the de-inking process vary substantially, it might be anticipated that non-linearities are expressed in the data. Additionally, seasonal variation might be present due to the long sampling campaign. Thus, it is of interest to apply GIFI-PLS to explore possible non-linearities.

20.5.2 Reference PLS model

PLS was utilized to correlate the 54 process variables to the COD load. A four-component PLS model showing the statistics $R^2X = 0.64$, $R^2Y = 0.62$, and $Q^2Y = 0.57$ was obtained. Considering the substantial composition variation in the recycled paper and the full-scale nature of the dataset, these values are excellent. The next two plots (Figures 20.11 and 20.12) show the scores of the first component and the loadings. The most influential input variables are found in the upper right-hand region of the loading plot.

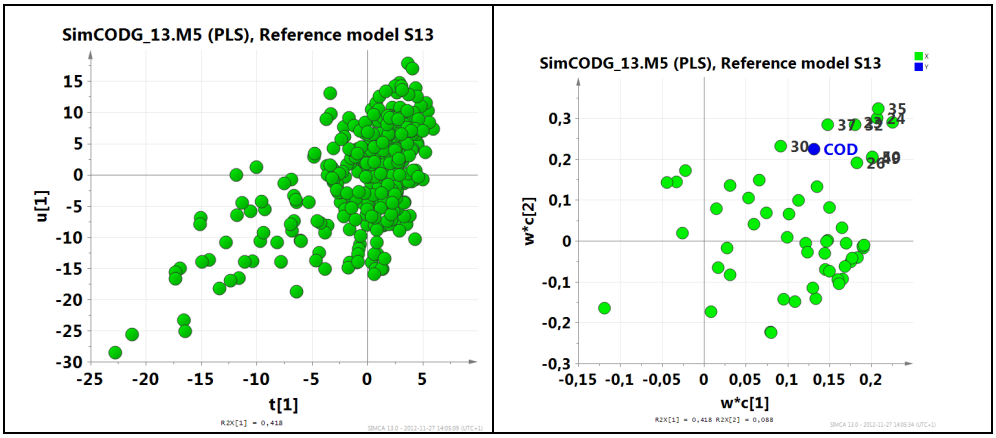


Figure 20.11: (left) PLS t_1/u_1 score plot of the conventional PLS model. The first component accounts for 36% of the sum of squares of COD. Figure 20.12: (right) PLS weight plot for the reference model. The position of the COD load, variable number 1, is given in blue color. The most important process variables are variables 23, 24, 26, 32, 35, 37, 49 and 50.

Moreover, the line plot of DModX against time order displayed in Figure 20.13 shows quite a few outliers. The critical distance (Dcrit) at the 0.05 level is shown as the dotted line. With 384 observations it is expected that around 20 (0.05 * 384 = 19.2) observations should be found outside Dcrit. In this case, however, as many as 57 time points display DModX-values greater than Dcrit. Finally, Figure 20.14 gives the relationship between observed and fitted COD values for the reference model. There is no doubt that the process variables carry important information for modeling and predicting COD load.

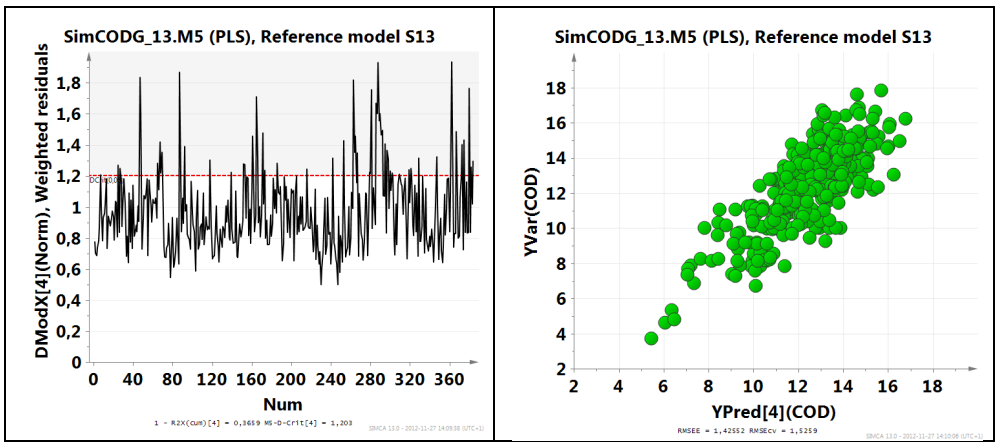


Figure 20.13: (left) The observation's distance to the model for the reference model. The critical distance at the 0.05 level is depicted by the dotted red line. Figure 20.14: (right) Relationship between measured and fitted COD values for the reference model.

20.5.3 PLS modeling using original and binned process variables

In the next stage, the 54 X-variables were binned using 5 intervals, thus producing 54*5 = 270 new categorical variables. We then tested the combined block of 324 X-variables (54 + 270) and acquired a four-component PLS model. This model had $R^2X = 0.27$, $R^2Y = 0.75$, and $Q^2Y = 0.60$, and is therefore more promising than the reference model.

It is interesting to compare the amount of explained Y-variation of the first components of the reference model (using 54 descriptors) and the PLS model founded on 324 descriptors. The first component of the reference model accounts for 36% of the COD variation (see inner relation plotted in Figure 20.11). The corresponding value for the new PLS model is 40%, i.e., an increase of 4% (see inner relation plotted in Figure 20.15). Additionally, the observed COD versus fitted COD plot provided in Figure 20.16 looks nicer than before (compare with Figure 20.14). Further comparison between the two models in terms of DModX indicates that the number of outliers has not changed, therefore no additional residual plots are provided.

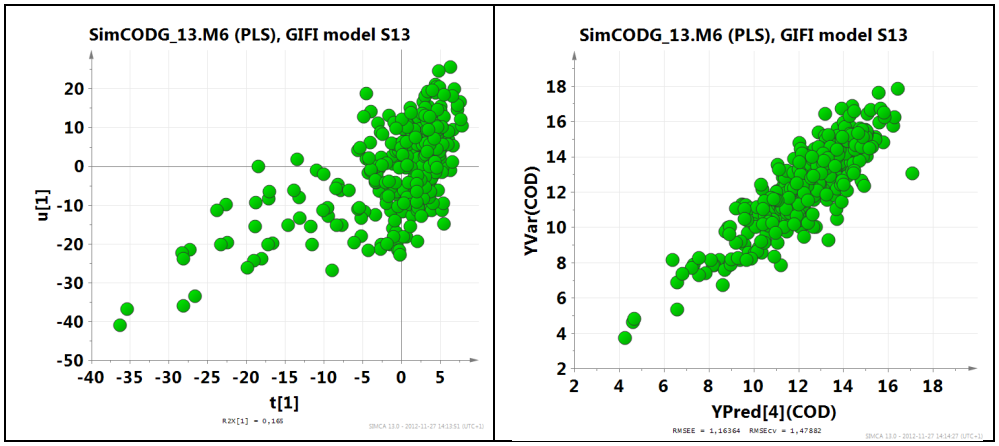


Figure 20.15. (left) PLS t_1/u_1 score plot of the improved PLS model. The first component reflects 40% of the sum of squares of COD. Figure 20.16: (right) Relationship between measured and fitted COD values for the PLS model established on both original and binned X-variables.

20.5.4 Where lies the non-linearity?

The resulting 324 regression coefficients will indicate which X-variables are linearly related to the COD load, and which are not. However, we will not enter into any detailed mechanistic interpretation; our ambition is just to exemplify how these linked sets of categorical variables may aid in the detection of subtle phenomena (e.g., non-linearity,

discontinuity, bimodality, etc.) perhaps not recognized during the interpretation of the reference PLS model.

The example variable highlighted below is variable x30, a process condition variable. This variable is not amongst the most influential as far as the model terms of continuous variables are concerned, it is ranked as number 14-15 depending on which model parameter is considered. However, the variable *does* rank very high when considering the categorical variables (Figures 20.17 and 20.18). In fact, the categorical variable representing the third interval (bin), v30b3, has the largest (negative) regression coefficient of all 324 variables.

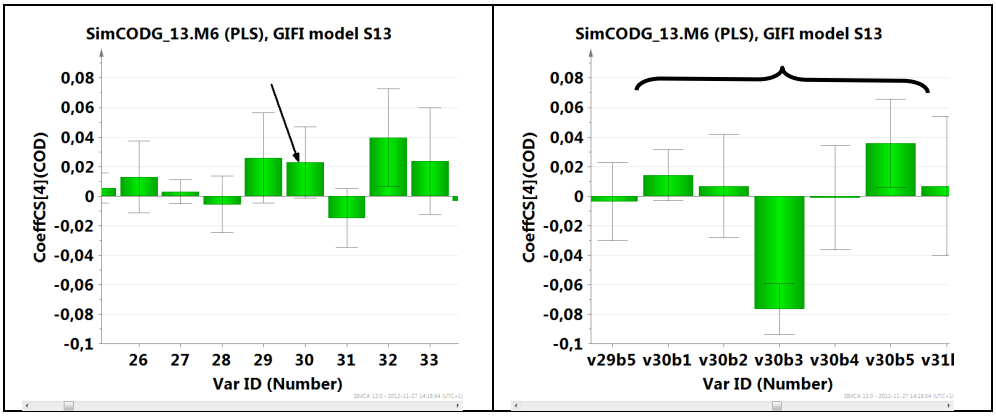


Figure 20.17: (left) Regression coefficients for some of the continuous variables of the PLS model fitted to the complete set of 324 X-variables. The arrow indicates the coefficient belonging to variable x30. Figure 20.18: (right) Regression coefficients related to some of the categorical variables. The brace shows the location of the five binned variables of x30. The profile of these five regression coefficients indicate deviation from linearity in the relation between x30 and COD.

The brace in Figure 20.18 points to the five expanded terms of x30. When referencing this coefficient profile against the patterns plotted in Figure 20.6(d), it is seen that the relation between x30 and COD is not linear. Additionally, we remember from Figure 20.17 that this relation is comparatively weak. That this is indeed the case is corroborated by the raw data plotted in Figure 20.19. We can see that for low and high values of x30, dense clusters of data points appear, whereas for intermediate values of x30 scattered data points are encountered. The relationship between COD and v30b3 is displayed in Figure 20.20.

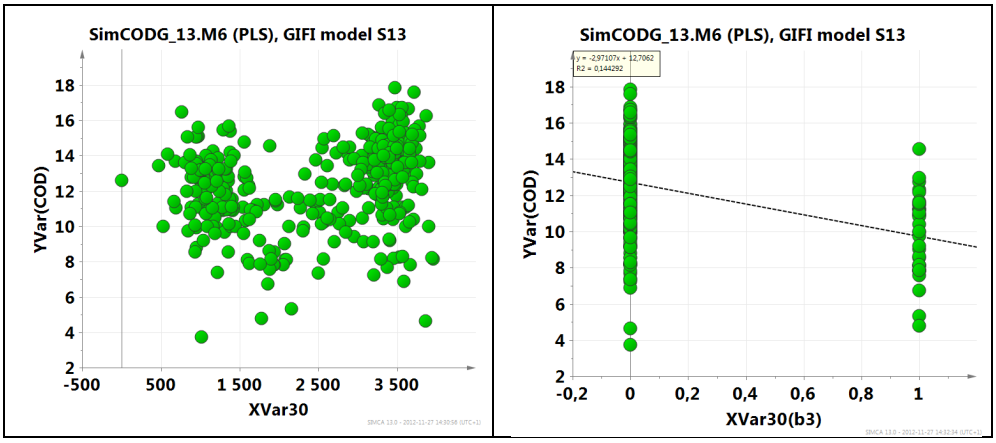


Figure 20.19: (left) Raw data plot of COD against x30. Figure 20.20: (right) Raw data plot of COD versus v30b3.

In order to better understand the transition from Figure 20.19 to Figure 20.20, one can think of the former plot as being vertically sliced into five equally large portions, where each slice represents one interval (bin). Because the numerical scale of x30 is roughly 1 – 4000, the first bin would be 1 – 800, the second 801 – 1600, etc. Hence, the interesting third bin would contain all those points of Figure 20.19 falling in the interval 1601 – 2400. It is

immediately evident that in this interval the average COD value is much lower than in all the other four intervals. Hence the negative regression coefficient of v30b3, as is also indicated by the raw data plot shown in Figure 20.20.

Similar searches for non-linearity, discontinuity and other anomalies, may now be conducted for all the other process variables. An inspection of the regression coefficients indicates that variables x6, x7, x8, x16, x34, x39, x46, and x50 exhibit the strongest deviations from linear x/y-relations. Figures 20.21 and 20.22 display the diagnosis of non-linearity for x8, i.e., (i) a small coefficient for the parent variable and (ii) a curved coefficient profile for the variables x08b1 – x08b5.

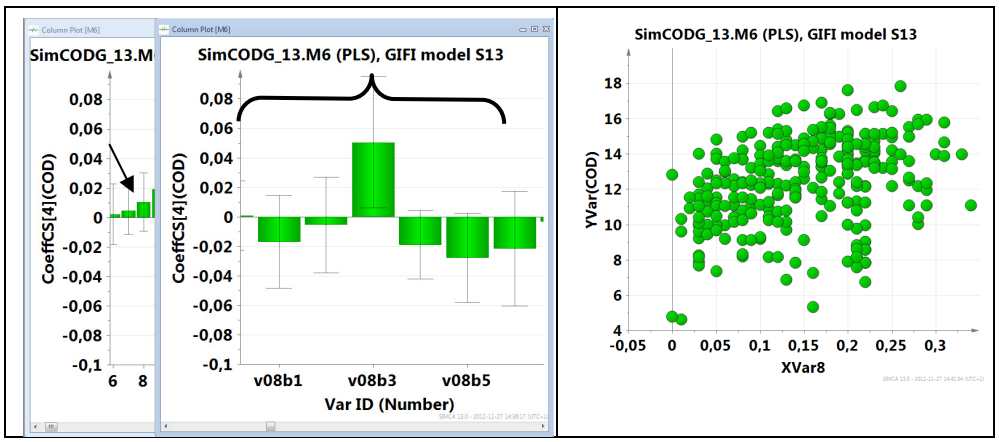


Figure 20.21: (left) Regression coefficients of continuous x8 (arrow) and expanded x8 (brace). The pattern for v08b1 – v08b5 is indicative of non-linearity. Figure 20.22: (right) The curved relationship between COD and x8.

20.5.5 Summary

In summary, we infer that the collected process variables are sufficiently informative to predict COD load at a level of $Q^2 \approx 0.6$. This value was verified with external validation on independent test data, though these results are not quoted in this section. Hence, the results obtained are very satisfying, especially when considering the full-scale nature of the data and the gross variations that may occur in the composition of the starting material (recycled paper). It is clear that there are non-linearities between some of the process variables (e.g. x8 and x30) and the COD load. A more thorough investigation of this dataset, in which the time series properties are also exploited, is found in the literature [Eriksson, *et al.*, 2001].

20.6 Questions for Chapter 20

1. In which fields are non-linear problems common?
2. What are the different degrees of non-linearity that may be discerned?
3. List some approaches to handling non-linearity!
4. Which PLS diagnostic tool is especially powerful in detecting non-linearity?
5. What is the main disadvantage of most tools for non-linear regression modeling?
6. What is PLS-RSM? INLR? Spline-PLS?
7. What is the rationale behind GIFI-PLS?
8. What are the main steps of GIFI-PLS?
9. What does the regression coefficient profile look like for a set of linked and "binned" variables in the case of linearity between the X-variable in question and Y? In the case of non-linearity?

20.7 Summary and discussion

When a system or process is subjected to large changes, it often appears non-linear. In the present context, this means that the relation between X and Y becomes non-linear. Also the relations between the X-variables may become non-linear, as well as the relations between the Y's. Numerous approaches have been published for the simultaneous estimation of the X-scores (t) and the parameters in a given type of non-linear inner relation.

Mild non-linearity may occasionally be treated by a linear PLS model simply by adding one or two components to compensate for the non-linear features encountered. However, a carefully developed model based on some non-linear regression – e.g. PLS based – method will also function well.

In general, great caution must be exercised in any type of non-linear modeling, including that of non-linear PLS. Since non-linear models are much more adaptable and flexible than linear models, they may easily fit outliers, noise, separate clusters, and the like, which results in very low predictive power of the model. A prudent use of cross-validation to avoid too many terms in the model is strongly recommended. Also, PLS score plots (t_a/u_a-type) provide diagnostics for the presence of non-linearities. A non-linear model is warranted only when deviations from linearity are seen in these score plots.

As mentioned in the introduction, there are many methods available for addressing non-linear problems, but in this chapter much emphasis has been placed on PLS-based approaches to estimate non-linearity. Non-linear PLS modeling may be accomplished e.g. (i) by transformation of X and/or Y, (ii) by expansion of the X-matrix (INLR and PLS-RSM), (iii) by formation of a non-linear inner relation (Spline-PLS, classical non-linear PLS), or (iv) by binning and expanding of X-variables into sets of 1/0 dummy variables (GIFI-PLS). A major aim of this chapter has been to introduce GIFI-PLS.

The use of binning of quantitative X-variables into categorical variables, coupled with the subsequent expansion of the resulting discrete variables into sets of 1/0 dummy variables, creates a flexible system that readily captures deviations from linearity. By interpreting regression coefficients of binned and expanded variables, it is possible to pinpoint where the non-linearity occurs, that is, whether all X-variables or just a few relate non-linearly to the response(s) modeled. Also, in applications involving very many variables, the binning of all variables seems a time-consuming choice. One possibility which might be useful is the binning of latent variables, such as a set of PCA or PLS score vectors. Here, however, more experience is needed.

Moreover, it is noted that the number of bins cannot be made too large, as a sufficient number of data points should ideally fall into each bin. Clearly there is a trade-off between the number of bins and the precision of the resulting model coefficients. To have some stability in the coefficients, it is recommended to have at least ten observations in each bin [Berglund, *et al.*, 2001]. This means that GIFI-ing does not apply well to small datasets with $N < 30$, since three bins is the smallest meaningful non-linear representation. In the ELASTASE ($N = 89$) and SimCODM ($N = 384$) examples we used five bins. Additionally, Berglund and coworkers observed "having more bins than around ten seems to provide little gain, even in the case of large datasets, but again this may depend on the application. More experience is needed in this area before any firm guidance can be given."

With only a few X-variables, it is also possible to employ a varying number of intervals for each X-variable, especially if prior information is available. If X-variables turn out to be linear in their relationship with Y, the original variables can then be re-instated. This will minimize the number of noise-variables added to the model. It is also important to study the distribution of the X-variables, since long-tailed distributions may lead to problems with the binning. Often this can be corrected by transforming the X-variables before the binning.

As noted by Berglund and coworkers, the original GIFI approach – where each X-variable is divided into bins and each bin is represented by a discrete 1/0 variable – does not allow predictions outside the interval spanned by the training set. This because an observation with a much higher value than any in the training set will be placed in the highest bin regardless of its real value. As a way of overcoming this limitation, Berglund and coworkers

proposed four modifications, GIFI I – GIVI IV, of the methodology. In these four approaches the basic binning step is the same, but the differences occur in the expansion step. Instead of using ones and zeroes in the new variables, the original variable values are used, sometimes with and sometimes without thresholding. For more details please consult the literature [Berglund, *et al.*, 2001].

21 Hierarchical cluster analysis, HCA

21.1 Objective

This chapter presents hierarchical cluster analysis (HCA), which is a tool for exploratory data analysis. HCA is useful to search for both subtle and strong group formations among observations (matrix rows) and variables (matrix columns). In HCA, distances between the observations (or variables) are calculated and compared. When distances between observations are relatively small, this implies that the observations are similar in some respect. When distances between observations are larger, this implies that the observations are markedly different. This chapter discusses HCA and exemplifies how it can be used to search for cluster formations in a process dataset arising from a six month measurement campaign of a continuous process. The visualization of HCA results in terms of a dendrogram plot is illustrated. As will be shown, the dendrogram plot can be connected to scores (or loadings) plots thereby making it possible to overview the clusters that exist and also understand their taxonomy. The basic principles underpinning the dendrogram plot are given.

21.2 Introduction

Cluster analysis, or simply clustering, is the task of assigning a set of observations (samples, objects, cases, items, …) into groups (called clusters) in such a way that the observations in the same cluster are more similar in some sense to each other than to those in other clusters. In other words, the objective of cluster analysis is to organize data points into smaller and more homogenous groups where the within-group variability is minimized and the between-group variability is maximized [Acar, *et al.*, 2000; Liao, 2005]. Cluster analysis is a main task of exploratory data mining, and a challenging issue for multivariate data analysis in many fields, including process modeling, multivariate calibration, pattern recognition, image analysis, and 'omics'. When existing clusters are few, "tight" and well separated in multivariate space, like small islands with lots of water in between, a technique like PCA will work well to detect and visualize groupings in the data (see Section 10.4.3). However, when clusters are many, less distinct, possibly also partly overlapping, and are captured by higher-order latent variables, there is a risk that groupings in a dataset will not be fully appreciated using PCA. It is in this latter respect that dedicated cluster analysis algorithms become relevant, i.e., to detect both subtle and prominent clustering in high-dimensional and complicated datasets.

It is important to realize that with the term *cluster analysis* we do not refer to one specific algorithm, but rather to the general data analytical task to be solved. Cluster analysis can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the multivariate space, intervals or particular statistical distributions. The appropriate clustering algorithm and parameter settings depend on the individual dataset and the intended use of the resulting clusters. Typical adjustable parameters include values such as the distance metric to use, the linkage function to use, the number of expected clusters, etc. Due to the influence of such adjustable parameters, cluster analysis cannot be regarded as an automatic task, but rather it is an

iterative process of knowledge discovery or interactive modeling (e.g. outlier removal) that involves trial and error. It may often be necessary to iteratively modify any occurring pre-processing and adjustable parameters until the final result is obtained.

21.2.1 What is a dendrogram?

Results of cluster analysis are conveniently visualized using a dendrogram. The principles behind the dendrogram and how to interpret it are discussed below (Figure 21.1).

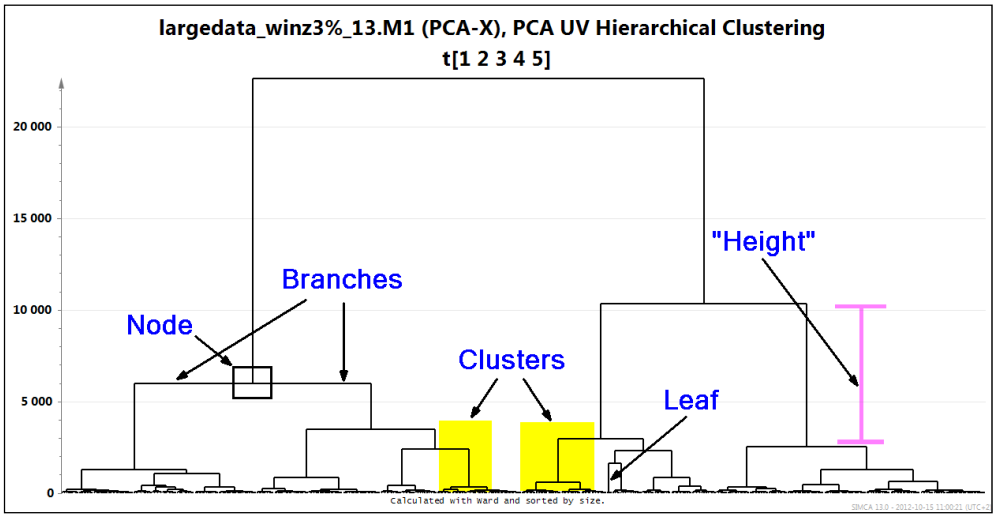


Figure 21.1: An example dendrogram. The vertical scale is a similarity measure, with similarity increasing as the numerical value gets closer to zero.

Originating from the Greek words of dendron ("tree") and gramma ("drawing"), the dendrogram is a tree diagram frequently used to illustrate the connectivity of clusters produced by cluster analysis. Its vertical extension corresponds to a similarity/dissimilarity measure with similarity increasing as the numerical value of the scale approaches zero. The following are the main constituents and concepts of a dendrogram:

- Leaves or leaf nodes: The bottom row of leaves, or leaf nodes, represent the individual observations (or variables) and thus there will be as many leaves as there are observations (or variables) in the training set.
- Clusters: Data point clusters are formed by joining individual observations or existing clusters with the join point referred to as a node.
- Node: The leaves or leaf nodes mark the entry of each single observation into the dendrogram. The remaining nodes, higher up in the tree structure, then indicate how individual observations first are merged to form small clusters, then how smaller clusters are jointed to form medium-size clusters, and finally how medium-size clusters are merged into successively larger clusters until only one super-cluster with all observations exists at the top of the tree.
- Branches: At each node there is a right and a left branch indicating the formation of two sub-clusters from the current cluster.
- Height or distance: The height of each node can be thought of as the distance value between the right and left branch sub-clusters. In general, the distance between merged clusters is monotonically increasing with the level of the merger, and the height of each node in the plot is therefore proportional to the value of the intergroup dissimilarity between its two daughter nodes.

21.2.2 Bottom-up and top-down approaches

Broadly speaking, the approaches to cluster analysis can be categorized as two types: agglomerative and divisive. Also known as the bottom-up approach, in the agglomerative approach, the point of departure is treating each observation as its own cluster. Pairs of clusters are then merged as one moves up in the hierarchy among clusters until only one super-cluster containing all observations remains. Conversely, the divisive clustering is initialized by treating all observations as one cluster, and cluster splits are then performed as one moves down the hierarchy towards smaller cluster sizes and increasing similarity. Consequently, the divisive approach is sometimes termed top-down clustering.

The SIMCA software supports two approaches for accomplishing clustering, i.e., hierarchical cluster analysis (HCA) and PLS-Tree. In this chapter we describe and exemplify HCA whereas Chapter 22 deals with PLS-Trees. HCA is a bottom-up method and the PLS-Tree is a top-down technique. HCA can segment the data table both row-wise and column-wise, whereas the PLS-Tree focuses on row-wise (observation-wise) clustering.

21.3 HCA in the SIMCA software

21.3.1 Recommended procedure involves scores or loadings

Hierarchical cluster analysis can be used to search for clusters among observations and variables. The point of departure is a distance matrix -- based on Euclidean distances -- containing all the pair-wise distances between the observations or variables. For very large datasets the calculations of such distance matrices, and updates thereof after cluster mergers, is computationally very taxing. Hence, the recommended approach in the SIMCA software is to apply HCA to the parameters of a multivariate model, i.e., the scores or loadings depending on what are the target cases (observations or variables). An example of HCA as applied to PCA score vectors was given in [Rännar and Andersson, 2010]. In principle, HCA can be applied to any parameter of a multivariate model, e.g., residual vectors, but it is envisioned that scores and loadings are the vectors that are used most frequently.

21.3.2 Choice of linkage criterion

In order to decide which clusters should be combined, a linkage criterion is needed. A linkage criterion quantifies the dissimilarity between clusters as a function of the pair-wise distances of observations in the various groups. The SIMCA software supports two linkage criteria denoted Ward and Single linkage (Figure 21.2), of which the former is the default choice.

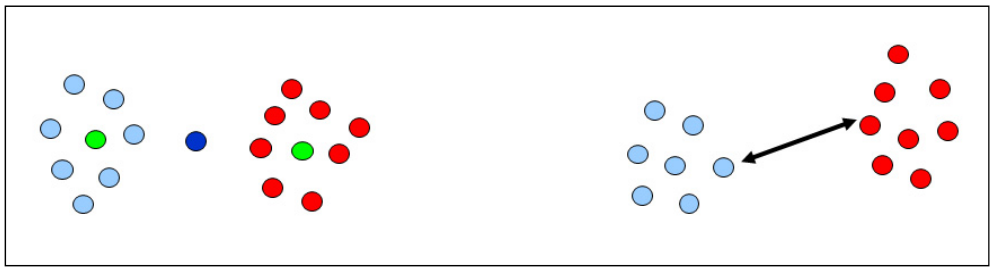


Figure 21.2: An illustration of the two linkage criteria supported by the SIMCA software. To the left is seen the Ward criterion and to the right the Single linkage criterion. In single linkage clustering, the distance between a pair of clusters corresponds to the Euclidean distance between the two observations (one in each cluster) closest to each other. In the Ward clustering, the difference in the sum of squares around the mean of each cluster (symbolized by the green points) before and after merging two clusters is considered. The candidate singleton (one-point cluster) will enter into the cluster resulting in the smallest increase in the Ward´s variance function.

Using the Ward linkage criterion, or Ward variance function, the clustering minimizes the total within-cluster variance. Thus, at each step, the pair of clusters is sought that leads to minimum increase in the total within-cluster variance after merging. The variance function used in SIMCA is the classical variance function of Ward, which calculates the difference in the sum of squares around the mean of each cluster before and after merging two clusters

[Ward, 1963]. Being biased towards globular clusters, Ward's linkage criterion is less susceptible to noise and outliers compared with other linkage criteria.

In the single linkage clustering, the distance between a pair of clusters is the Euclidean distance between the two observations (one in each cluster) closest to each other. This type of clustering works best when clusters are elongated and not globular. A drawback of single linkage is that clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other. This effect may arise when there are outliers present that perturb the specification of what are the clusters.

Alternatives to single linkage exist and these are based on longest or average distances between samples in two clusters. The latter two approaches, denoted Complete linkage and Average linkage, are not supported by the SIMCA software, however.

21.3.3 Facilitated interpretation by connecting the dendrogram plot to other plots

The groups or clusters found by a HCA can be colored in the resulting dendrogram. In order to facilitate interpretation of the clustering results, any relevant and open plot is then colored accordingly. Similarly, the points inside each cluster are listed together in the Marked Items dockable window. In this way, it is handy to get hold of the members of any cluster formed by HCA. Figure 21.3 is an example of an observation-wise clustering where interesting clusters are detected in the dendrogram and interpreted by means of the concomitant score scatter plot.

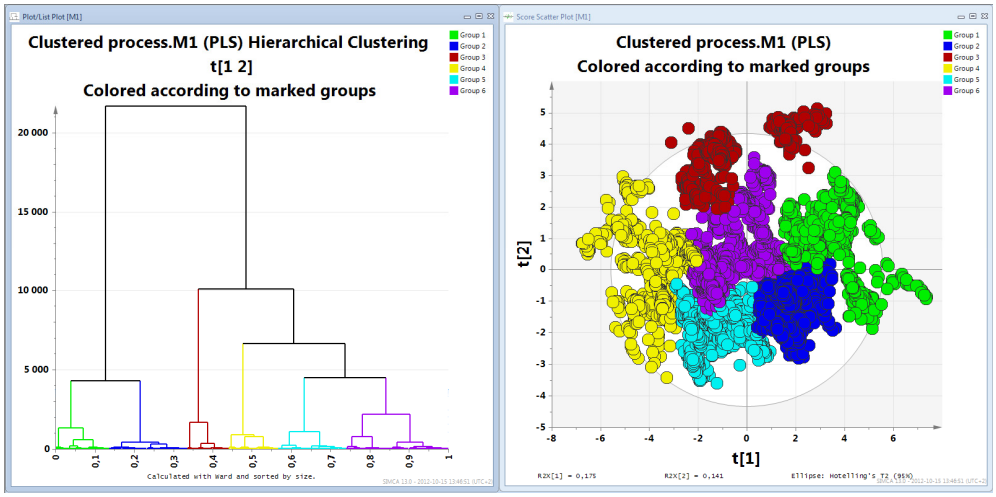


Figure 21.3: Open dendrogram and other plots are connected to make the interpretation easier. The example relates to row-wise clustering and how the distribution in score space of the observations is quickly understood.

21.4: An example: A clustered process

21.4.1 Introduction to the dataset

We shall here relate the clustering results of a tutorial dataset having an obvious clustering. The objective is to briefly illustrate the main data analytical steps of clustering, as well as to demonstrate how easy it is to interpret the dendrogram plot if coupled to a score scatter plot. The dataset arises from an industrial production site where data were logged during a six month sampling campaign. The process is a continuous process, which is known to undergo substantial changes due to changing production conditions. This variability arises due to changes in machine speed in a critical step of the process, as well as changes in the type of product being manufactured. The dataset contains 7250 observations, i.e., time points at which process measurements were made, and 31 variables.

21.4.2 Outlier detection and removal

As discussed above, the performance and reliability of hierarchical cluster analysis can be negatively impacted by the presence of outliers. As a consequence, an initial PCA model was computed, which, according to cross-validation, had six significant principal components. Figure 21.4 shows the Hotelling's T2 and DModX control charts tiled together. Recall that control charts for multivariate statistical process control (MSPC) purposes were introduced and illustrated in Chapter 15. The presence of some gross outliers is immediately clear. Both control charts detect six strong outliers. These have to be removed prior to the hierarchical cluster analysis.

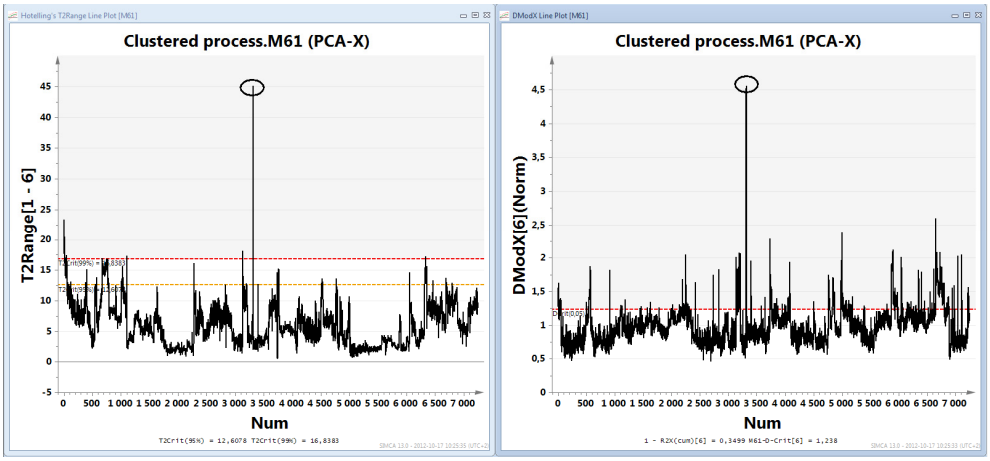


Figure 21.4: MSPC control charts tiled together. Both charts highlight six strongly deviating observations (time points close in time) that are marked by the ellipse.

21.4.3 Updated PCA overview model

An updated PCA model, devoid of the six outliers, was fitted to the adjusted 7244 by 31 data-table. Again six principal components resulted, with the performance measures R2X = 0.65 and Q2X = 0.38. The scores of the two first components are visualized in Figure 21.5. As shown by Figure 21.5, there is a clustered behavior of the process, and a relatively large cluster that can tentatively be identified by "eye-balling" is framed. The distribution and relevance of this tentative cluster can be checked more stringently using HCA.

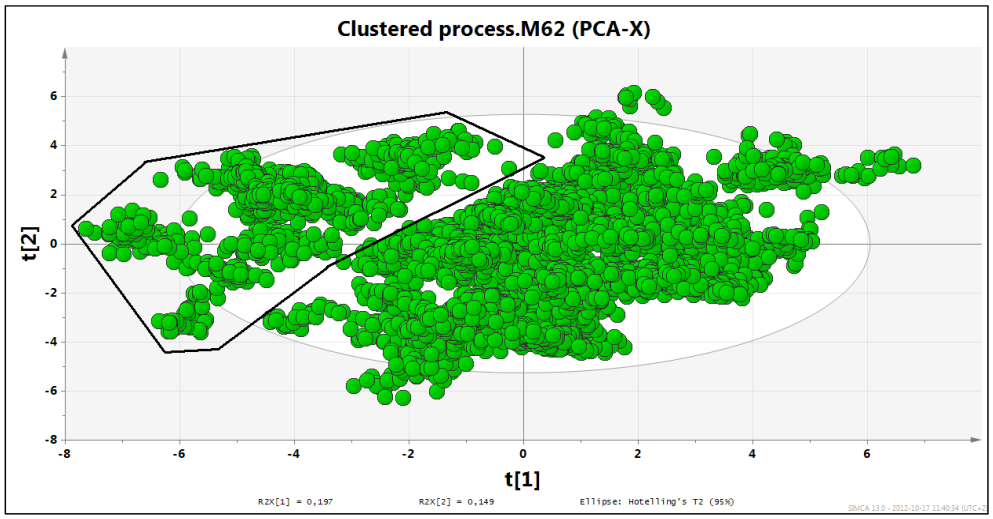


Figure 21.5: PCA score plot of the clustered process. An obvious grouping in the first two scores is indicated by the frame. The relevance of this cluster across the full model, the six principal components, can be checked using hierarchical cluster analysis.

21.4.4 HCA in the six principal component score vectors

To get a full run-down of the clustering situation, we applied hierarchical cluster analysis to the six principal component score vectors using the default Ward linkage criterion. The resulting dendrogram is plotted in Figure 21.6

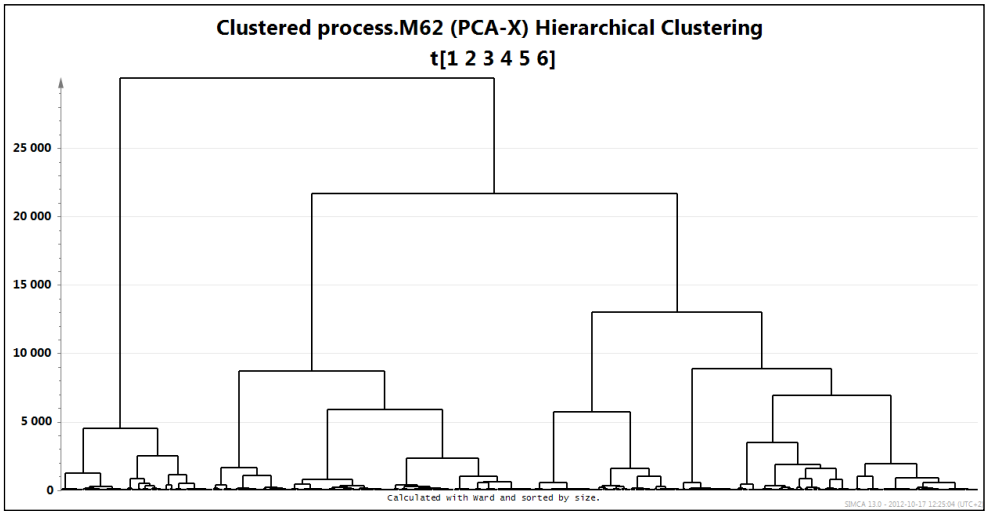


Figure 21.6: The dendrogram arising from HCA of the six score vectors using the Ward linkage criterion. According to the dendrogram, the dataset contains at least three major clusters.

The dendrogram following the Ward linkage criterion suggests a structure to the process dataset comprising at least three main clusters. Figure 21.7 displays the distribution of the three clusters. The conclusion is that the HCA corroborates the existence of the cluster that was framed in Figure 21.5.

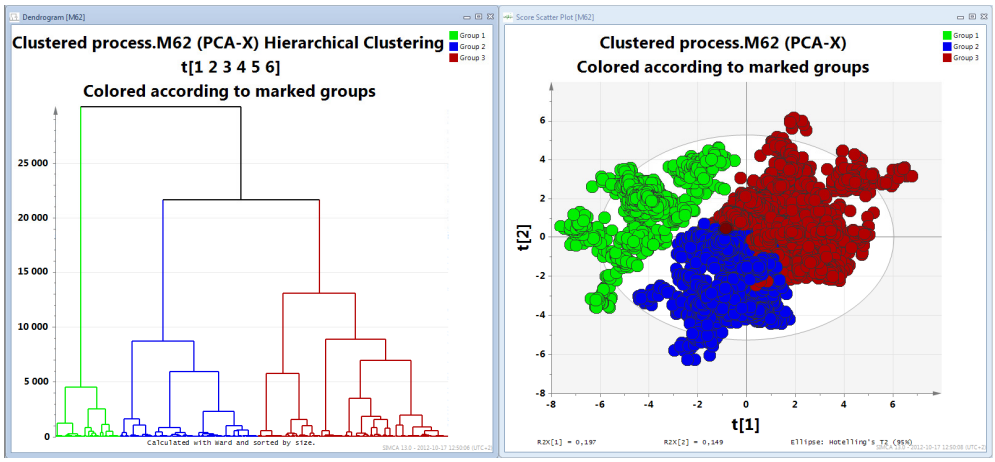


Figure 21.7: Since the dendrogram plot and the scatter plot of the scores are connected, the visualization of how the clusters are distributed in score space is straightforward. The uniqueness of the green cluster cannot be mistaken.

21.4.5 Evaluation of the resulting cluster pattern

Since the examined process is a continuous process, there is a time order among the observations. This relationship can be exploited by making a line plot of, for instance, the first score vector, see Figure 21.8. Such a line plot makes it completely clear that the green cluster in Figure 21.7 is related to the last two months of the sampling campaign. This points to the fact that the covariance structure among the 31 process variables is markedly different during the last two months compared with the first four months.

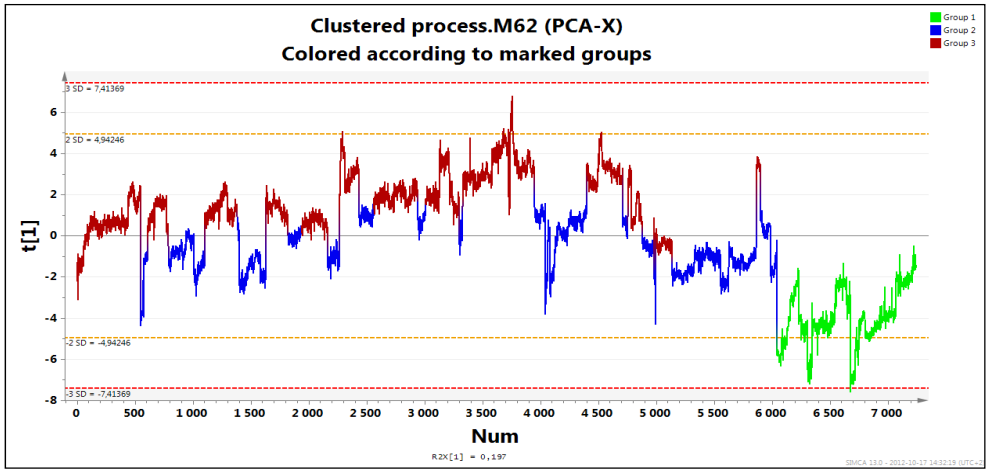


Figure 21.8: Line plot of the first PCA score vector, where the horizontal order is the time order. The total time span is six months. The green cluster that was suggested by the PCA in Figure 21.5 and corroborated by the HCA in Figure 21.7 is very localized in time and occurs during the last two months of the sampling campaign.

In order to interpret which process variables are the driving forces behind the clustering of the green cluster, it is useful to create a score contribution plot. Recall that the principles for the contribution plot were outlined in Chapter 9. The left-hand part of Figure 21.9 provides such a contribution plot, pointing to the process variables number 2, 4, 10 and 26 as the primary determinants for the late-stage process deviation.

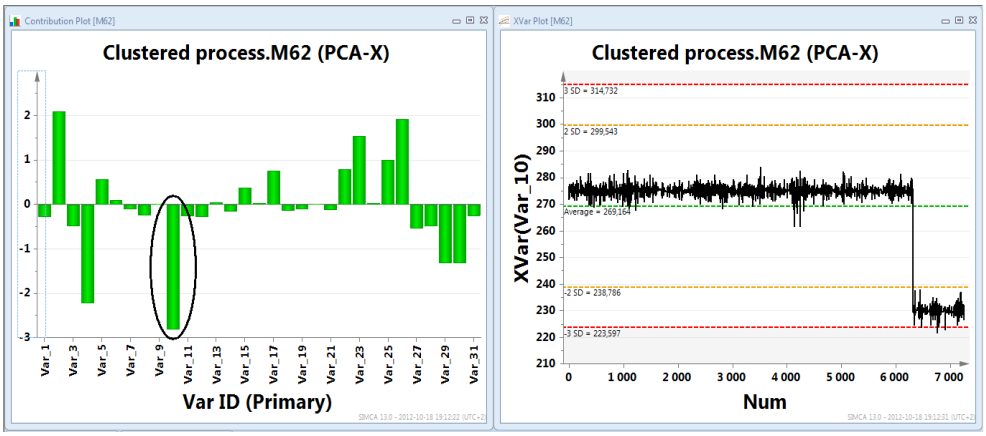


Figure 21.9: Contribution plot (scores mode) indicating which variables express the dissimilarity of the green cluster. One of the variables that is uncovered is process variable number 10. The line plot in the right-hand part of the figure shows a sharp drop in this variable during the latter part of the sampling campaign.

21.4.6 Summary of the example

The case study discussed above indicates how hierarchical cluster analysis can be used to make a more formal account of clusters that are (i) either not known beforehand or (ii) perhaps indicated by an initial PCA model computed for overview purposes. The number of clusters detected and their shape will depend on the clustering algorithm and linkage criterion used. This is further discussed below, see Section 21.6. The important point to remember is that whenever clusters among observations are detected, this signals that the covariance structure among the measured variables is not stable, and changes when moving from one cluster to another.

It is crucial that the clusters that are detected are thoroughly interpreted to justify their postulation and relevance. One way to accomplish this is through contribution plotting to discover the "culprit" variables that are governing the clustering. In the current process example it was found that four process variables had significantly different trajectories towards the end of the measurement period, i.e., in the last two months out of the six months that were monitored.

Further testing of the clustering may involve using the SIMCA methodology, i.e., disjoint principal components models for the various clusters that have been detected. How to apply the SIMCA methodology to compute local class models was discussed in Chapter 10. The SIMCA approach was pursued in the current example with respect to the green cluster and it was concluded that this cluster was very different. A local PCA model fitted to the green cluster samples diagnosed, in terms of a predicted DModX chart, the remaining process time points as not belonging to that model (results not shown).

Another aspect that should not be overlooked is that the outcome of hierarchical cluster analysis is sensitive to the presence of outliers. The extent to which the clustering is affected by outliers cannot be anticipated in advance, however. This depends on the number of outliers, their position in multivariate space, their severity, and which clustering algorithm and linkage criterion are used. The hierarchical cluster analysis, therefore, involves trial and error, and hence our emphasis above that it is crucial to carefully interpret the clusters that result.

The main conclusion of the analysis of the clustered process is that the clusters that are encountered by the PCA model are relevant and can be corroborated by the HCA model.

21.5 Questions for Chapter 21

1. What is hierarchical cluster analysis? When is it useful?
2. How can HCA complement a PCA model?
3. What is a dendrogram? How can it be interpreted?
4. Which are the main constituents/elements of a dendrogram?
5. What does bottom-up clustering mean? Top-down clustering?
6. Why is it often necessary to apply HCA to parameters of a multivariate model rather than original measurements?
7. How can HCA be used to segment a data-table?

21.6 Summary and discussion

As noted in Section 10.8.1, a difficult question in multivariate classification and discriminant modeling is that of specifying the classes. The definition of classes must be done with caution and correspond to a real and observable group similarity. The similarity within a class must be greater than the between-class similarity. Thus, an essential objective of the initial exploratory data analysis is to reveal whether a sound class division has been accomplished or not. An obstacle in classification and discriminant modeling is that of large inhomogeneities in classes believed to contain similar observations. Although multivariate methods may classify prediction set observations correctly despite inhomogeneities in the classes of the training data, the results are likely very difficult to interpret and utilize since the class concept is violated (it is strongly linked to the idea of intra-class similarity). In SIMCA, PLS-DA and OPLS-DA, inhomogeneities and other anomalies in the training data are indicated by weak models and/or deviating observations in scores and model residuals.

Hierarchical cluster analysis may play a central role in the definition of new clusters/classes, as well as corroborating the existence of clusters/classes that have been detected by PCA or perhaps are anticipated based on prior knowledge or experience. From the viewpoint of cluster recognition, one may therefore regard HCA as an exploratory tool that is complementary in nature to PCA. It is complementary in the sense that, while PCA helps to uncover cluster that are distinct, HCA will search for clusters that are subtle. HCA will of course also find clusters that are very pronounced. HCA makes no assumptions on the number of classes. In fact, any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. Thus, in many cases, HCA is a useful starting point for more in depth data analysis using local multivariate projection models (PCA, SIMCA, PLS-DA & OPLS-DA).

The outcome of HCA will depend on the distance metric used and the linkage criterion that is selected. The SIMCA software only works by considering Euclidean distances, so that´s not an issue; however, there is a choice between the linkage function due to Ward and the single linkage approach. HCA based on the Ward linkage criterion is known to work best when clusters are globular. Experience shows that it tends to form clusters high up in the dendrogram, i.e., at a high level of dissimilarity, that contain fairly many members and extend in a balanced fashion across the latent variables constituting the basis for the clustering. By contrast, the single linkage clustering tends to form clusters that are elongated and sometimes also with a bit of discontinuity attached to them. The single linkage clustering is known to be sensitive to outliers. It tends to produce clusters at a high level of dissimilarity in the dendrogram that have markedly different numbers of members and span the underlying latent variable space in an unbalanced and irregular fashion.

Thus, in summary, the recommended procedure to HCA involves Ward's variance function (lower sensitivity to outliers) and utilizing the parameters of a multivariate model as input (for reasons of computational speed). HCA is most often used for clustering of observations, but its ability to cluster correlated variables should not be forgotten.

22 PLS-Trees

22.1 Objective

This chapter presents a top-down clustering approach denoted PLS-Trees, which is based on a set of nested PLS models. This approach is analogous to classification and regression trees (CART), but uses the scores of PLS regression models as the basis for splitting the clusters, instead of the individual X-variables. The split of one cluster into two is made along the sorted first X-score vector (t_1) of a PLS model of the cluster. The position of the split is selected according to the minimization of an objective function, a variance function. Compared with conventional hierarchical cluster analysis, the clustering based on PLS-Trees is much faster. The PLS-Tree method is illustrated using a data mining example involving pharmaceutical quantitative structure-activity relationship (QSAR) data.

22.2 Introduction

22.2.1 General considerations

The large amounts of data created in all parts of science and technology are presenting interesting challenges. The management, analysis and visualization of large datasets – often called data mining – are different in several aspects from the dealing with "ordinary" datasets of moderate size, with, say, up to a few hundred observations and variables.

A common problem is the inhomogeneity of large datasets, where relationships between variables are different for different subgroups (clusters) of the data. If not identified and acted upon, this situation results in more or less trivial models dominated by group differences. Using such models for predictions gives low quality results, with the predicted values being near the averages of the closest group (cluster). The only known remedy is to segment the data into smaller, fairly homogeneous and balanced groups (clusters). Here outliers can be considered special groups with just one or a few observations. This segmentation will, if successful, allow the estimation of relationships among both variables (columns of a data table) and observations (rows of a data table) with good predictive power.

Hence, data mining, and other analyses of large and complex datasets, should involve one or more clustering approaches in the early stages of the data analysis, preferably combined with an embedded multivariate regression step (i.e. PLS, OPLS, O2PLS, or similar). Such a multivariate clustering would model a large dataset with potentially many and collinear variables as a set of models (clusters). The complexity of the observational direction is reduced from N observations to G clusters, and the complexity in the variable direction is reduced from K variables to A_g scores (latent variables), with A_g typically being different for the G clusters (index g).

Since large datasets usually have missing data, noise, and other complications, the clustering approaches need to be able to handle such complications. Another desirable property for a clustering approach is computational speed. A "top-down" approach starting from the whole dataset, and successively dividing the data into smaller parts, is usually faster than a "bottom-up" approach, starting with clusters of size 1, combining the two closest to a cluster of size 2, and continuing the merging until all observations have been combined to a single, final cluster. The latter approach needs to keep track of all inter-observation distances as well as observation-cluster and cluster-cluster distances

(dissimilarities). This becomes computationally costly when the number of observations, N, exceeds a couple of thousand.

22.2.2 Latent variable-based approaches to clustering

The idea of adopting a latent variable-based approach to hierarchical clustering is not new. One attempt was presented by Böcker *et al.*, who introduced a hierarchical clustering algorithm called NIPALSTREE [Böcker, *et al.*, 2006]. This is a PCA-based procedure, which at each level of the tree projects a dataset onto the first principal component. Thereafter, the dataset is sorted according to this one dimension and split close to the median position. In a final step, the clustering is recursively applied on the resulting subsets until the maximal distance between cluster members exceeds a user-defined threshold. Because NIPALSTREE uses PCA, it will be relatively slow on large datasets. Furthermore, only the X-block is taken into consideration in the clustering, and any Y-data are not influencing the results.

A slightly different approach was taken by Barros and Rutledge in their development of the so called DiPLS_Cluster method [Barros and Rutledge, 2004]. The DiPLS_Cluster method starts with a random binary (0/1) Y-vector, i.e., two random groups. In an iterative fashion the Y-values are replaced by new values predicted by the PLS-DA algorithm using a one-component model. After convergence, the predicted Y-data are rounded off to the nearest integer (0 and 1) and used for cluster assignment. Hence, this approach works only with the X-data although it uses PLS for accounting in its internal calculations. A modified approach, called Generalized PLS_Cluster has also been reported [Bouveresse, *et al.*, 2007]. This approach allows any reasonably small number of clusters in each node of the dendrogram, instead of just two.

To overcome some deficiencies of the DiPLS_Cluster approach, Thomas and coworkers published a method named DiPCA_Cluster [Thomas, *et al.*, 2008], which they claim to be "an optimal alternative to DiPLS_Cluster". The basic difference between the two approaches is that Thomas *et al.* strive to define a stringent objective function for the clustering, which leads to a PCA based approach, using only the X-part of the data.

22.2.3 A fast, robust and flexible approach

Our objective with this chapter is to review a top-down hierarchical clustering approach denoted PLS-Trees [Eriksson, *et al.*, 2009; Byrne and Wold, 2011]. The approach is based on PLS to enable the responses (Y) to influence the clustering; i.e., the Y-data are explicitly used as a part of the splitting criterion.

The main result of the approach is a hierarchical tree (dendrogram) of connected PLS models, i.e., the PLS-Tree. An example PLS-Tree is displayed in Figure 22.1. Dendrograms produced by the HCA (see Chapter 21) and PLS-Tree methods have similar building blocks and lay-out, and the only interpretational difference relates to the meaning of the vertical scale. While in HCA the vertical scale of a dendrogram represents a cluster similarity/dissimilarity measure, the vertical scale in a PLS-Tree dendrogram reflects cluster size or cluster proportionality. The cluster size scale runs between 0 and 1, where 1 designates the single global cluster containing all observations (100%).

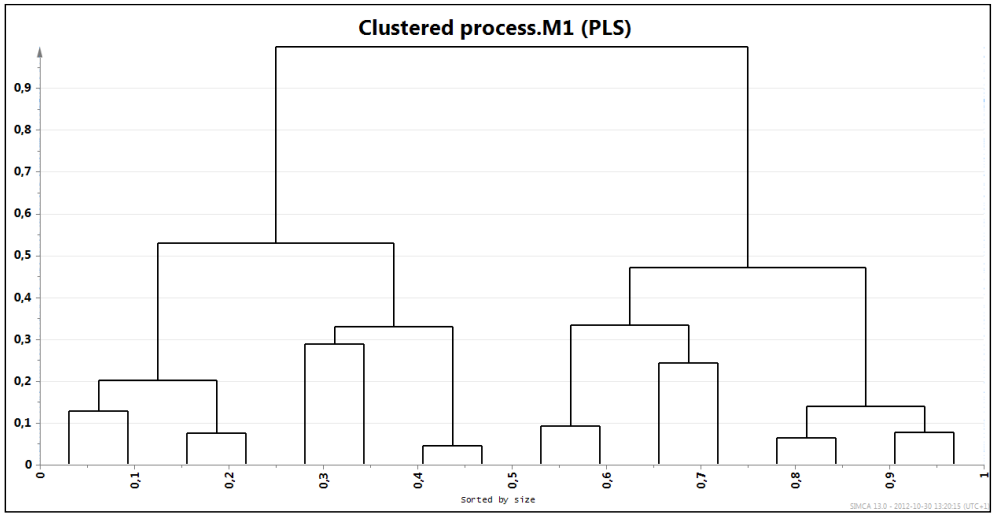


Figure 22.1: An example PLS-Tree. The vertical scale reflects cluster size or cluster proportionality, with the value of 1 given to the single super cluster at the top containing 100% of the observations.

Conceptually, the PLS-Tree method is analogous to classification and regression trees (CART) [Breiman, *et al.*, 1984], but it uses the scores of PLS regression models as the basis for splitting the clusters, instead of individual X-variables. As discussed below, the PLS-Tree approach is fast, robust and flexible.

22.3 Theory of the PLS-Tree method

The point of departure of a PLS-Tree is to start with a global PLS model of the whole dataset and then split the data along the sorted first score vector, t_1. Of course, prior to the calculations, the user must specify what is X and Y, and any scalings and other related pre-processing steps. For clarity, we note that the first score vector, t_1, is obtained as $t_1 = Xw$, where $w = X^Ty/\left\|X^Ty\right\|^2$. Recall that the basics of PLS and OPLS were discussed in Chapters 4 and 5. It should also be pointed out that t_1 is different in each model, and hence lower layer PLS models cluster along different directions than the upper layer PLS models.

Subsequently, in the PLS-Tree clustering algorithm, the position of the split along t_1 is selected by minimizing an objective function. The objective function is a variance function consisting of three terms, (a) the variance of the X-score t_1, (b) the variance of Y, and (c) a penalty term encouraging a balanced split with approximately equal numbers of observations in each resulting cluster. More specifically, after sorting all observations from maximum to minimum score value along t_1, we look to identify the "cut" point on t_1 that divides X and Y in two parts, 1 and 2, such that the following expression is minimized:

$$B*\frac{(N_1-N_2)^2}{(N_1+N_2)^2}+(1-B)*\left(A*\frac{V_{y1}+V_{y2}}{V_y}+(1-A)*\frac{V_{t1}+V_{t2}}{V_t}\right)$$

In the expression above, V_{Y1} is the variance of Y for group/cluster 1 above the cut, V_{Y2} is the variance of Y for group/cluster 2 below the cut, V_Y is the variance of the entire Y-block, V_{t1} is the variance of the t-vector above the cut, V_{t2} is the variance of the t-vector below the cut, V_t is the variance of the entire t-vector, N_1 is the number of observations above the cut, and N_2 is the number of observations below the cut. The values of the user entered parameters A and B lie between 0 and 1. The first parameter, A, sets the balance between the score t_1 and Y; the closer to zero, the more weight is attributed to the score t_1. The second parameter, B, takes into account the group size of the resulting clusters; the closer to zero, the less important it becomes to have similar group sizes in the dendrogram. In addition, a smallest resulting sub-cluster size – N_{min} = min (N_1, N_2) – is specified to disallow a solution with many very small clusters even in the case where the parameter B is zero or very small. The default in the SIMCA software is $N_{min} = 5$.

On closer inspection, the output value of the objective function above is a weighted combination of three terms constituting the variance expression. Figure 22.2 shows, for a PLS root model in a typical PLS-Tree, line plots of these three terms and the total variance function against the sorted t_1 score vector of the root model. In summary, this means that a division along t1 is sought that minimizes the variation within a group and hence maximizes the differences between groups in t1 and Y, also taking the difference in group size into consideration.

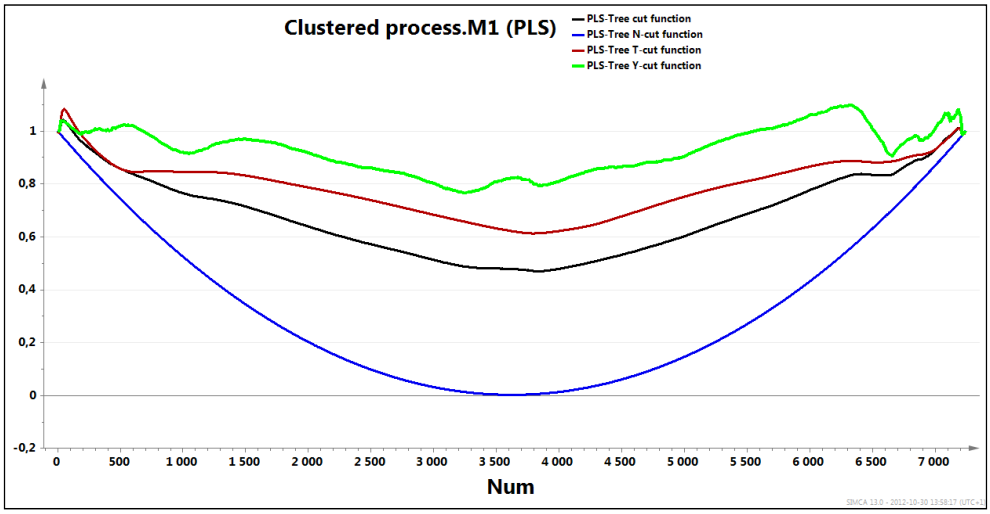


Figure 22.2: Line plot showing the evolution of the PLS-Tree variance function and its three components against the index of the sorted t_1 score vector (called "Num") of the root PLS model of a typical PLS-Tree. The black line represents the total variance function, the red line its t-variance term, the green line its Y-variance component, and the blue line its subgroup size part. Being a weighted combination of the red, green and blue lines, the shape of the black line will change depending on the settings of the adjustable parameters A and B. The cut position will be the point on the black line where the minimal value is encountered. In the current case, the minimal value is found at "Num" 3824 and hence the N=7250 observations of the PLS model are split into two subgroups containing 3834 and 3416 observations.

Cross-validation is used to terminate the branches of the tree, and to determine the number of components of each cluster PLS model. The user may also specify the desired number of layers in the PLS-Tree and the smallest number of observations inside a cluster. The maximum depth, D, of the PLS-Tree – i.e., the maximum number of layers in the tree – is typically set to 4 (default) or 5. The number of models making up the tree will be $2^{D+1} - 1$, e.g., 31 PLS models with D = 4, and 63 models with D = 5. We note that the clustering speed in the exploratory phase is enhanced considerably by restricting all PLS models to one single component. Later, in the interpretation phase, models of interest can be extended to include all CV-significant components (autofitting).

22.4 An example: DataMining-QSAR

22.4.1 Background to the dataset

The current example concerns a data-mining application of QSAR data in a pharmaceutical database. The objective is to investigate whether the PLS-Tree approach may indicate useful clusterings in the historical dataset pointing to local and stable structure-activity relationships (SAR). Data-mining of historical data represents a highly useful approach in the early phases of pharmaceutical research projects, when usually large numbers of compounds are also experimentally investigated for their biological activity with respect to a certain target. The current dataset involves 5000+ compounds mapped by 500+ physico-chemical descriptors. The single y-variable, the endpoint, reflects metabolic degradation of the compounds. All data are coded for proprietary reasons.

22.4.2 Data-mining via top-down clustering

A PLS-Tree with depth D = 5 was computed and the calculations took less than a minute. The resulting dendrogram is seen in Figure 22.3.

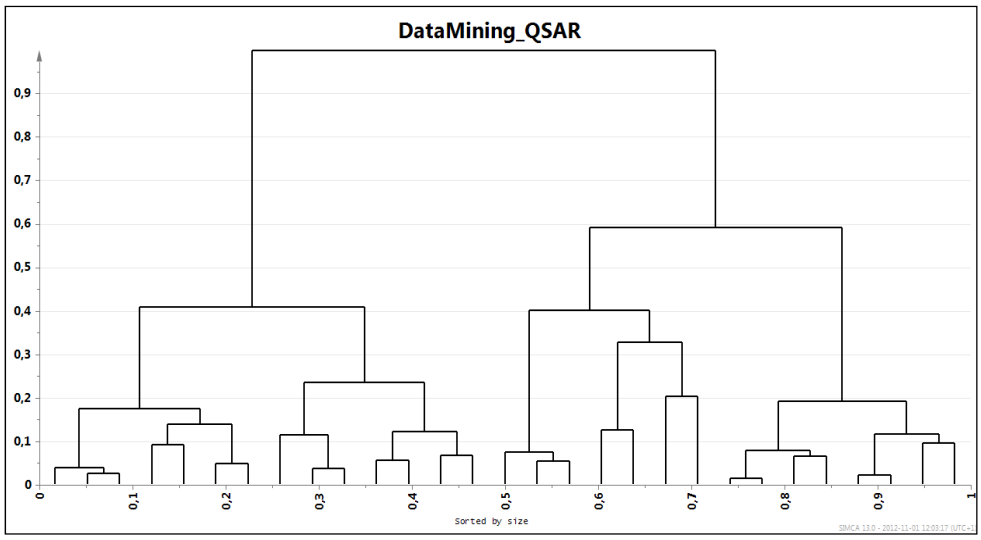


Figure 22.3: The PLS-Tree of the DataMining-QSAR example. It comprises 55 clusters and PLS models.

As discussed above in Section 22.3, a PLS-Tree of depth D = 5 may cover up to 63 PLS models, corresponding to 63 clusters of varying size. In the current case, however, the result was 55 models. This means that in some cluster splits there was no significant PLS model to use as a basis for the splitting criterion.

Figure 22.3 shows that the first split of the QSAR data results in two clusters containing roughly 40% and 60% of the compounds respectively. In the next "layer" of the PLS-Tree, the larger 60% cluster is divided in two clusters encompassing approximately 20% and 40% of the observations, whereas the smaller 40% cluster is partitioned in two clusters containing about 18% and 22% of the compounds.

For every cluster that is detected a PLS model can be fitted. The relevance of each PLS model and underlying cluster of compounds can be evaluated by considering statistics like N (number of observations), A (number of components), R2X, R2Y, Q2Y, average Y and standard deviation of Y. The SIMCA software provides this type of information in an overview window, whereby the most "promising" clusters and models can readily be identified.

22.4.3 Zooming-in on stable structure-activity relationships

Once an interesting PLS model with good statistics has been detected, the taxonomy -- "position" in the dendrogram -- of the underlying cluster is easily found by interactively marking the observations of said cluster in e.g. a scatter plot of the PLS scores or of the observed versus predicted Y (see Figure 22.4). Figure 22.4 demonstrates how the taxonomy of a cluster having 117 members (compounds) can be understood.

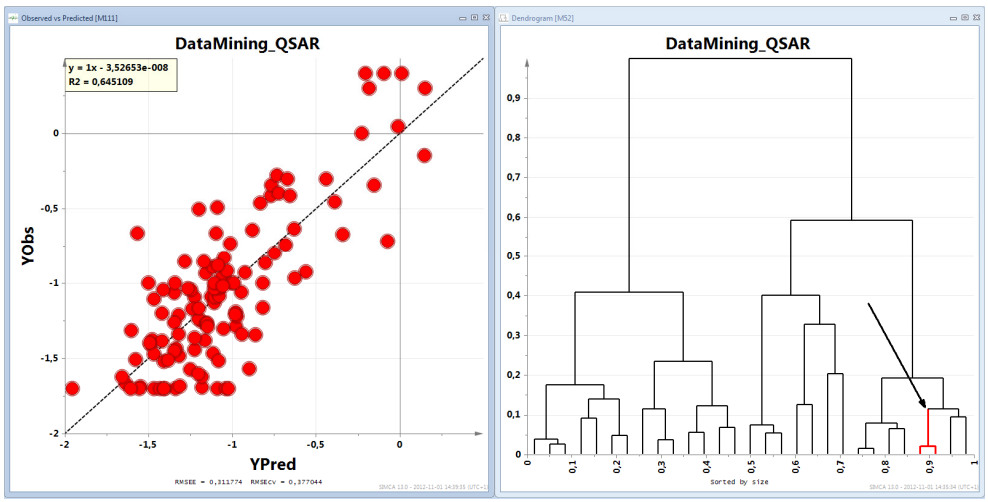


Figure 22.4: In a data-mining application using top-down clustering, the relationship of an interesting cluster to other clusters in the PLS-Tree can be understood by interactively marking the observations of the cluster in question. This marking is conveniently accomplished in e.g. a plot of observed Y versus predicted Y.

The relationship between the observed Y and the predicted Y in Figure 22.4 points to a highly useful QSAR model with good statistics. The explained variance of Y is 64.5% and the predicted variance of Y is 47% using three PLS components. Figure 22.5 shows the Validate plot obtained after 500 response permutations. The outcome of the response permutation test points to a strongly significant local QSAR model. The performance statistics of the local QSAR model can be compared with those of the global reference model covering all 5000+ compounds, which are R2Y = 48% and Q2Y = 43% accomplished using eight latent variables. Hence, the local QSAR model provides more detailed knowledge of how the chemical and pharmacological properties are related, and does so using much fewer PLS components.

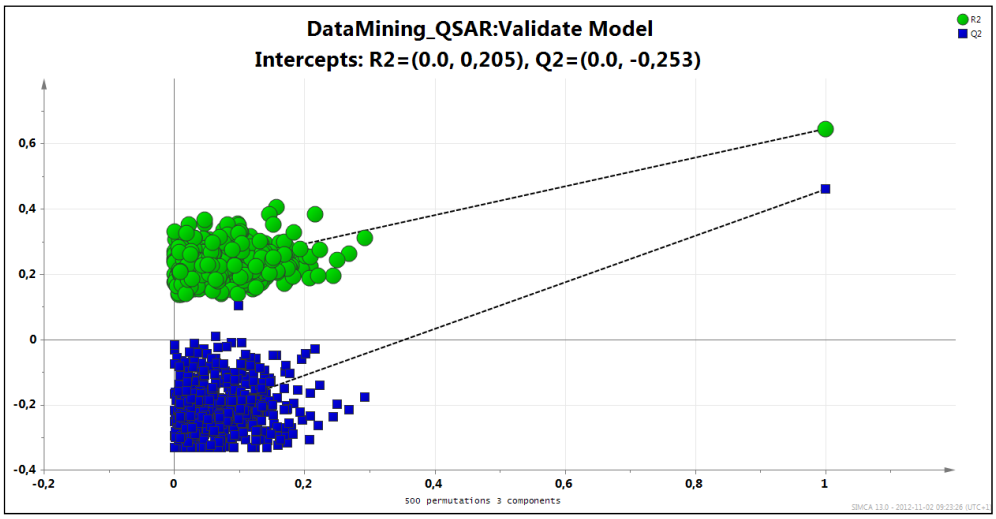


Figure 22.5: Response permutation plot of the local QSAR model pointing to a well-founded model.

Interestingly, the cluster of compounds constituting the basis for the local QSAR model is not the only useful cluster formation. In fact, in the PLS-Tree of Figure 22.4 there are three additional clusters that encode local QSAR models that have even higher model performance statistics than the local model discussed here. All these local models can be interpreted and refined the usual way by considering plots of scores, loadings and residuals. How to interpret PLS score plots, PLS loading plots and PLS residual plots was discussed in Chapter 4.

22.4.4 Summary of the example

The DataMining_QSAR example indicates how cluster analysis methods can be used to identify homogeneous data clusters encoding local models. These local models are very often useful and have better predictive power than models built with larger and heterogeneous data. Moreover, experience shows that such local models usually contain fewer components, since they are confronted with less latent structures than are global models. Hence, the local models are often easier to interpret.

The interpretation of an entire PLS-Tree is facilitated by the fact that each node has an associated ordinary PLS model. Hence the rich diagnostics of PLS such as plots of scores, loadings, VIP, coefficients and predictions are directly useful for this interpretation. In addition, subjecting the PLS-Tree clusters to supervised classification by means of OPLS-DA and/or SIMCA classification reveal such things as which variables are important for cluster separation, the separation of the clusters, and poorly classified observations.

The comparison of loading parameters of different models provides a measure of similarity/dissimilarity between clusters. For instance, scatter plots of the first loading vector of two models can be created to identify QSARs with similar and dissimilar loading profiles. Conceptually, such a plot is very similar to the SUS-plot discussed in Chapter 11.

22.5 Questions for Chapter 22

1. What is CART? When is it useful?
2. What is a PLS-Tree? When is it useful?
3. How can a PLS-Tree be used to segment a data table?
4. What is a dendrogram? How can it be interpreted?
5. What does top-down clustering mean? Bottom-up clustering?

22.6 Summary and discussion

We have here described a top-down clustering PLS-based approach, called PLS-Trees. One appealing feature is its speed, especially useful when clustering a large dataset. For example, in another data mining example involving 500,000+ observations and 100 variables it took less than ten minutes on a standard laptop to accomplish a complete run-down of the entire data using a PLS-Tree with depth D = 5. A comparative study on exactly the same dataset using hierarchical cluster analysis based on the first few principal components was too slow and no dendrogram was obtained.

As shown by way of the examples in [Eriksson, *et al.*, 2009; Byrne and Wold, 2011], the PLS-Tree is a versatile, flexible and transparent clustering tool that brings about a number of interpretative possibilities and improved models. The net result of many of the local models pertaining to the datasets discussed in these references were improved predictive power. Moreover, the PLS-Tree was also found to corroborate classes known or anticipated in one of the applications. This ability of the PLS-Tree to recognize and predict structured data is a very useful property.

PLS-Trees may be seen as an extension of classification- and regression trees (CART) in the same way as PLS and OPLS are extensions of multiple linear regression. CART and PLS-Trees apply to different kinds of data; the latter typically involves rank deficient data with many variables, the former few and fairly independent X-variables. In comparison with conventional regression approaches, PLS and OPLS can handle many, collinear and noisy X-variables and also handle moderate amounts of missing data in X and Y, if the latter is multivariate.

Regression trees may have some problems finding the optimal cut-off value of the currently investigated X-variable, as well as in highlighting the correct X-variable among a set of correlated X-variables. This has lead to the proposition of so called random forests, where several reminiscent and randomly generated trees based on similar cut-offs are pursued,

leading to improved predictive precision, albeit at the cost of increased computing time and interpretational complexity.

The PLS-Trees seem to have fewer problems with the instability of the cut-off values. This is probably due to two reasons. First, the smoothing of the current score by fitting sliding piecewise polynomials in a Savitzky-Golay fashion, relieves some of the burden of finding a stable cut-off value. Secondly, being based on score vectors (linear combinations of the individual X-variables) the selection of the right X-variables is less of a problem, since the score vector is a weighted average of all X-variables, where the "best" ones, highly correlated to the Y-data, are heavily weighted.

In applications where responses (Y-variables) are available, the PLS-Trees have the advantage over conventional "Y-independent" clustering in that the PLS-Trees improve the relation between Y and the clusters. This is analogous to how PLS may lead to a better Y-focused latent variable model than PCA. Interestingly, when Y comprises a discrete matrix with 1/0 columns corresponding to a number of predefined classes, the result is a PLS classification tree.

Furthermore, the two adjustable parameters A and B in the PLS-Tree algorithm provide a range of variants to address the current given problem:

- with A=1, the approach is fully focused on Y and hence similar to CART;
- with A=0, the approach is fully focused on t and the X-space and hence is more similar to classical cluster analysis including the NIPALSTREE and the methods of Ruthledge and coworkers;
- with A=0.3 a reasonable compromise is obtained between focus on Y and X;
- with B=0, the approach gives traditional cluster analysis which, in the beginning, often splits off many very small clusters and individual observations (outliers);
- with B=1, the approach forces the division of clusters into equal parts, which is a rather rigid approach;
- with B=0.3, a reasonable balance seems to be obtained.

The approach taken in [Eriksson, *et al.*, 2009] of running a formal experimental design in A and B, is recommended when knowledge about the data structure is questionable or nonexistent. This actually takes a step towards some kind of random forest approach.

The PLS-Tree is a versatile technology and several extensions of the basic methodology are conceivable. A very obvious extension involves searching along a combination of the first two (or three) X-scores. The inspection of cluster geometries in many applications makes it clear that clusters are often oriented in a fashion that is not parallel with the first X-score (t_1). By introducing a parameter c, valued between -1 and 1, and searching along combinations of t_1 and t_2, i.e., $[c\ t_2 + (1 - |c|)\ t_1]$, the whole plane of the first two components will be investigated instead of only the line along t_1. Moreover, by using t_2 of the first PLS model as a direction orthogonal to t_1 to further split clusters 1 and 2, and analogously for models in lower layers, the similarities between PLS-Trees and CART are strengthened, in the sense that tree-based methods like CART can easily be extended to splitting in multiple dimensions.

Furthermore, cluster models may also be based on OPLS/O2PLS and/or hierarchical models. Using OPLS/O2PLS for the cluster "mother model" makes the t_1 search direction in X-space more aligned with the Y-space, which may be desirable in many cases. We note that with ordinary PLS "mother models", the split point search is always along the X-score(s), even if the Y-variance comes into the search criterion. With many X-variables (and/or Y's), hierarchical models (see Chapter 19) provide an interesting alternative to variable selection, and facilitate the model interpretation. The PLS-Tree approach would also easily work with hierarchical PLS or OPLS/O2PLS models, providing an interesting alternative for datasets with many X-variables.

In data-mining of "big data", the reduction of the set of variables is often desirable. With large numbers of X-variables, PLS models may also have a sub-optimal predictive power. Eliminating variables that either are strongly correlated to the best predictors or uncorrelated

to the Y's may improve the predictive power of the model and simplifies the model interpretation. The approach of "Selective PLS" [Kettaneh-Wold, *et al.*, 1994]], later used under the name IVS-PLS for interactive variable selection in PLS [Lindgren, *et al.*, 1994], provides a simple but powerful way to variable selection based on keeping only variables with large PLS weights (w_k or w_k*). Either all variables with smaller algebraic values than a certain limit, say 75 % of the largest weight, or the, say, 75 % of the variables with the smallest algebraic weights would be excluded.

In the present context this variable selection would start with the top (mother) model and the next layer would start with the reduced set of variables reducing it further until the lowest layer would have the smallest numbers of variables. Alternatively, one can of course restart the variable selection for each cluster model, and optionally also use different thresholds for the variable selection for the different layers in the PLS-Tree.

In summary, the PLS-Tree method provides an interesting and rapid approach to data mining and clustering of large datasets. The PLS-Tree approach can, like any PLS model, handle multiple and collinear variables, even more numerous than the number of observations. Moderate amounts of missing data are automatically handled by the PLS-Tree NIPALS algorithm. An advantage with the PLS-Tree approach is its reliance on ordinary PLS models, making the interpretation of results based on the well known diagnostic tool-box of these methods simplified. PLS-Trees is not a "new" method isolated from other data analysis, but rather another extension of PLS into the complicated domain of data mining in large and complex datasets.

Appendix I: Model derivation, interpretation, and validation

AI.1 Objective

The objective of this appendix is to give an overview of the main steps in multivariate data analysis. Briefly, the main steps are (i) evaluation of raw data, (ii) model derivation and interpretation, and (iii) model validation and use. The appendix contains a flowchart for multivariate data analysis.

AI.2 Introduction

AI.2.1 Main steps of MVDA

The analysis of experimental and process data consists of three primary stages, i.e., (i) *evaluation and pre-processing of raw data*, (ii) *model derivation and interpretation*, and (iii) *model validation and use*. The first stage, *evaluation and pre-processing of raw data*, focuses on getting a general appraisal of regularities and peculiarities in the data. In most cases, raw experimental and process data can rarely be used just as they are in a data base. Often they need to be re-shaped into a format better suited for sound multivariate data analysis. For instance, outliers may have to be removed, or skewly distributed variables may warrant a transformation, or there may be a need for signal correction and compression in calibration modeling. If needed, data may also be scaled and centered during this phase. In summary, the main aim of the evaluation of raw data is to prepare the data in such a way as to facilitate the subsequent multivariate modeling with PCA, PLS, OPLS, and O2PLS.

The second stage, *model derivation and interpretation*, involves the actual calculation of the PCA, PLS, OPLS and O2PLS models. Recall that the concept of a model, and the theoretical foundation for multivariate modeling were addressed in Section 2.7. PCA is used for data overview, e.g., for detecting outliers, groups, and trends among observations, and for evaluating relationships among variables, and between observations and variables. PLS is used for linking predictors (X) and responses (Y) together, with the aim of predicting the latter from the former. OPLS has a similar scope to PLS but has the advantage of simplifying model interpretation. O2PLS is used for data integration of several blocks of data, by defining unique and joint variabilities in the data. In any kind of modeling, the model interpretation is crucial. Not only is it important to understand the implication of a model, sound interpretation allows one to undertake concrete action in accord with the model indications.

However, prior to using a model for forecasting future outcomes, it is of crucial importance to validate the model and its predictive power. This is the essence of the third stage, *model validation and use*. To facilitate this objective, there are a number of informative model diagnostic and validatory tools available, which will be reviewed in this appendix. As far as the model validation is concerned, we will give a dedicated discussion regarding some model validation tools. Furthermore, the use of a PLS model may also involve the interpretation of response contour and response surface plots, for deciding where the best point of a system or a process is situated.

When using a model for routine analysis in an analytical laboratory, or in an on-line situation in a manufacturing industry, it is important to recognize that a one-time validation of a model is never sufficient. This is because there might be time-dependent changes in a system or a process, which the model may not have been calibrated against. Such changes might be different operating conditions, fluctuating environmental and seasonal variables, as well as varying properties of starting materials, which jointly contribute to slow to moderate time variations. The only way to carry out long-term testing of the relevance of a model is to continuously withdraw samples for independent test sets. Whether every 10^{th}, 100^{th}, or 1000^{th} sample should be removed for testing of model validity is, however, a separate discussion and the approach to this problem varies from application to application.

AI.2.2 The iterative nature of MVDA

Because raw data often need to be pre-processed and because many steps are usually needed in perfecting a model (e.g., removal of outliers, addition/removal of model terms, variable transformation, etc.), multivariate data analysis is very much an iterative procedure. It is quite normal to work through several analysis cycles, using different ways of pre-treating raw data and involving different kinds of multivariate models, just to see how things are. Figure AI.1 illustrates the iterative nature of the multivariate data analysis cycle.

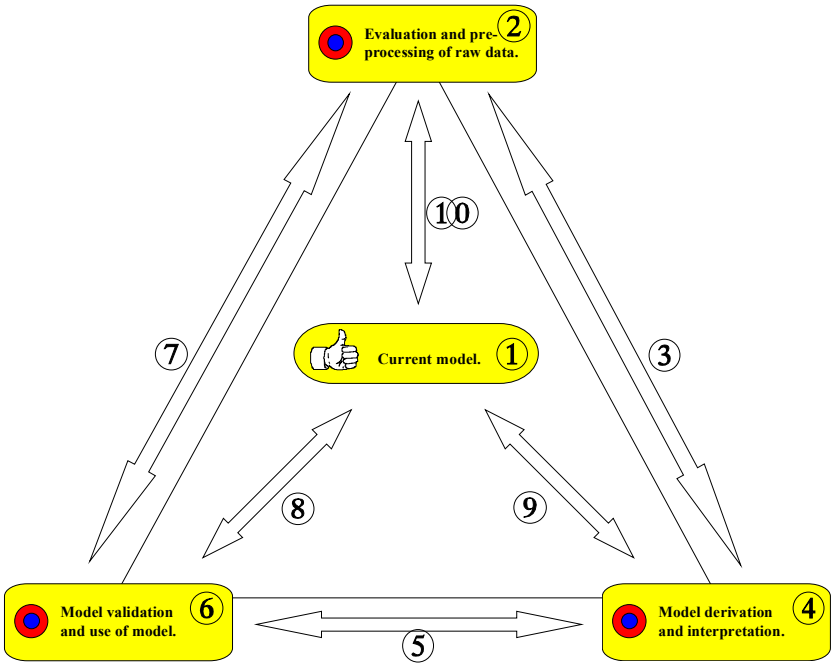


Figure AI.1: A schematic illustration of the multivariate data analysis cycle.

Figure AI.1 shows that there may be alternative ways of arriving at, and departing from, the current model (bullet 1). When starting a completely new project, it is advisable to spend some time evaluating and pre-processing the raw data (bullet 2). Often important insights about the quality of incoming data are obtained, which may simplify the subsequent multivariate modeling. After completing some of the data pre-treatment, a transition (bullet 3) to model derivation and interpretation takes place (bullet 4). Here, the multivariate model (PCA, PLS, OPLS, O2PLS) is computed and interpreted. It is important to use the richness of the diagnostic tools and model parameters (scores, loadings, weights, DModX, cross-validation, confidence intervals, etc.) to arrive at an optimal model.

Two main outcomes of the modeling are possible (bullet 4). On the one hand, it might be found that data modification is necessary to improve modeling conditions, e.g., it might be worthwhile to return to the previous step (bullet 2) and transform a variable or divide the observations into smaller groups, etc., followed by renewed modeling. On the other hand,

the pre-defined modeling conditions might be satisfied, whereupon a transition (bullet 5) to model validation and use of the model (bullet 6) normally occurs.

In the model validation step, a PLS or OPLS model is tested for its forecasting ability by computing predictions for an external test set of observations. A model may also be used to decide where to perform further experiments by inspecting plots of response contours or response surfaces. At this stage also, two main outcomes are possible. Either the predictions are found to be adequate for the purposes of the current investigation, or they are diagnosed as of insufficient quality.

If the predictive power of a model is low, one option is model refinement (check number of latent factors, eliminate variables, etc., bullet 4) to see whether prediction enhancement is possible. Here, however, care must be exercised so that model manipulation is not driven too far. The recommended procedure is to make a small adjustment to a model, then test the influence of this modification on an external test set. In fact, it is acceptable if this procedure is repeated two or three times, especially with DOE planned data. However, what should be strongly discouraged is a "systematic" approach in which many repetitive cycles of small model modifications and prediction testings are performed one after another on the same test set. The reason for this is that the final result is a conversion of the external test set into a secondary type training set.

Moreover, sometimes external samples are not predicted well, but may still be recognized as model members (e.g., they are inside DModX). In this situation, the new samples represent a new type of information, which should be used to update the training set. Consequently, the model needs to be re-calibrated with this new information. And, additionally, it might be necessary to revisit the raw data evaluation and pre-processing, and a transition (bullet 7) back to the first step (bullet 2) can therefore take place.

Whenever a model has been found adequate for its purposes, the next step corresponds to arranging its use (bullet 1) in an at-line and/or on-line real-world situation, for example in connection with SIMCA-online. Very often a model is determined to be ready-for-use after predictive validation (bullet 8), but sometimes a model may be regarded as "final" after diagnostic testing and evaluation (bullet 9). Finally, the drawing also contains a double-headed arrow (bullet 10) linking the current model (bullet 1) with evaluation and pre-processing of raw data (bullet 2). This arrow symbolizes the feed of newly pre-processed fresh data into an existing model, occurring, e.g., when executing a model on-line in a process industry.

In summary, MVDA is iterative, in fact highly iterative, and the derivation of an optimal model often encompasses many steps back and forth, as indicated by Figure AI.1. It should be clear from the above discussion that the derivation of a multivariate model is a handicraft work that requires training, experience, and persistence. However, there are many tools, diagnostics, and tests available to assist the user in the derivation of an optimal model. It is the aim of the next few sections to summarize these tools.

AI.3 Evaluation and pre-processing of raw data

AI.3.1 Main steps

On first obtaining a dataset it is very tempting to go ahead and fit a multivariate model just to see how things are. However, it is often beneficial firstly to consider evaluation and pre-processing of the raw data. A careful evaluation of raw data and intelligent pre-processing makes the data analysis much easier and smoother. The main steps here are:

- Inspection of univariate statistics
- Inspection of time series properties
- Trimming and winsorizing
- Transformations
- Expansions of X-matrix (use DOE first!!!)

- Scaling and centering
- Signal correction and compression

In earlier chapters, we have described the essentials of transformations and expansions (Chapter 12), scaling and centering (Chapters 3 and 13), signal correction and compression (Chapter 14, and elucidation of time series properties using auto- and cross-correlation functions (Chapter 17).

AI.3.2 What should we look for in univariate statistics?

As in any modeling, a good start is to examine the univariate statistics to understand, e.g., percentage of missing data, skewness of variables, minimum and maximum values, standard deviations, variances, and so on. In our experience, a joint assessment of (i) skewness of variables and (ii) ratio of minimum to maximum values is a very powerful indicator for when a transformation might be warranted. When the absolute skewness exceeds 1.5 or 2 and when the min/max ratio is smaller than 0.1, it is appropriate to check whether transformation of a certain variable is needed (see exemplification in Section 12.3). Also, it is important to know the proportion of missing data, as too many gaps may inhibit the development of good models.

AI.3.3 Efficacious plots

Once a suspect variable has been found its distributional properties can be examined in a histogram. Figures AI.2 and AI.3 exemplify the analytical power of histograms, for a variable prior to (Figure AI.2) and after (Figure AI.3) logarithmic transformation.

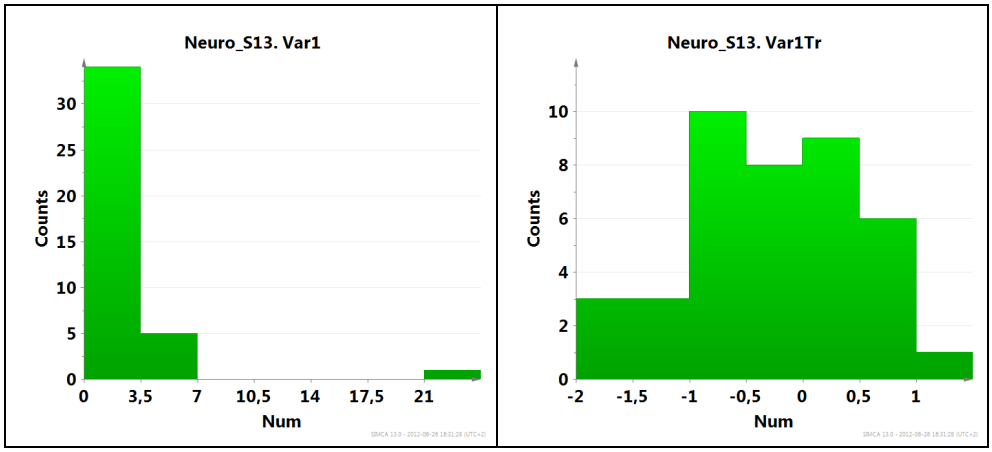


Figure AI.2: (left) Histogram of a non-transformed variable Var1. Figure AI.3: (right) Histogram of variable Var1 after log-transformation.

Another informative plot is a scatter plot of raw data to examine correlations among variables. The scatter plot can be used for investigating how two predictor variables are correlated, how a response and a predictor are associated, and what are the inter-relationships between pairs of response variables. Particularly with non-designed process datasets, scatter plots of raw data may help in deciding when expansions of X with interaction terms are appropriate. When two variables are by nature very well correlated (Figure AI.4) there is no point in trying to estimate their interaction, but when two variables have a rather good spread in a scatter plot (Figure AI.5) trying to include their interaction might be a fruitful alternative.

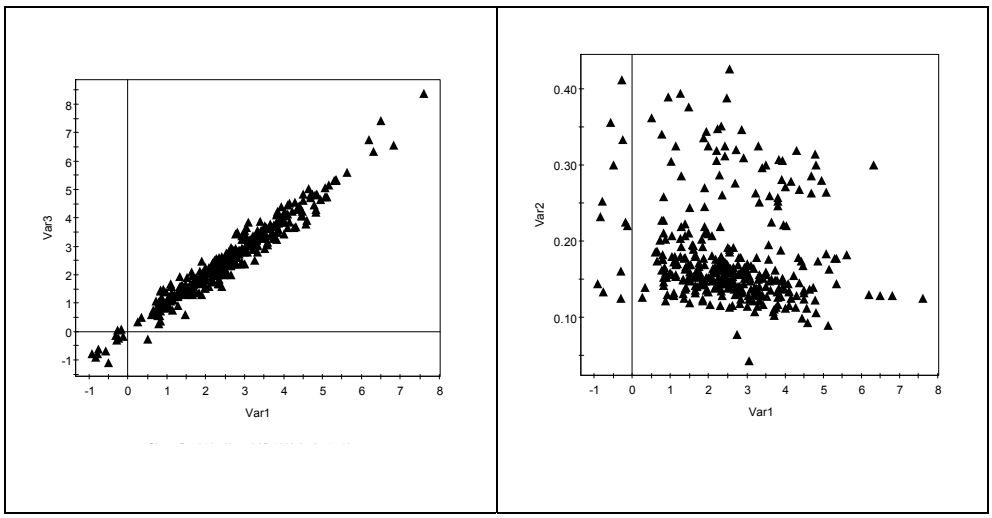


Figure AI.4: (left) Scatter plot of two correlated predictor variables in a process industry. Because of their strong relationship their interaction term (if any) cannot be estimated. Figure AI.5: (right) Scatter plot of two relatively uncorrelated predictor variables, whose interaction might tentatively be estimated.

AI.4 Model derivation and interpretation

AI.4.1 Main steps

Having evaluated and pre-processed the raw data, one may proceed to the next phase of model derivation and interpretation. Here, it is strongly recommended to initiate the modeling with a PCA of the entire dataset, regardless of whether disjoint PCA (SIMCA) or PLS/OPLS/O2PLS modeling is the final operative goal. The reason for this is that PCA gives an overview of the data, which is useful for directing further modeling efforts. Once the intrinsic properties of a dataset have been understood, a more refined and quantitative modeling with PLS/OPLS/O2PLS would usually be performed. The main steps here are:

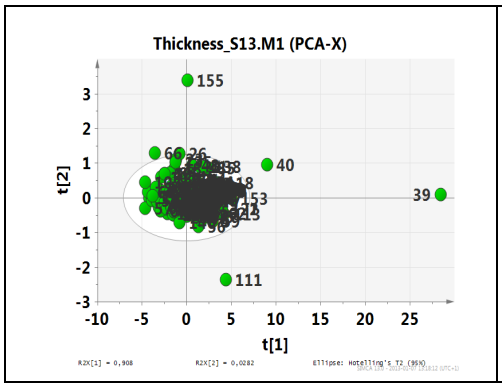
- PCA for data overview
- Interpretation of PCA model
- Refined modeling with PCA, PLS, OPLS or O2PLS
- Interpretation of refined model

In earlier chapters, we have described the essentials of PCA (Chapter 3), PLS (Chapter 4), OPLS (Chapter 5) and O2PLS (Chapter 6), and the tools they offer for model diagnosis and interpretation. Below, a condensed summary of the most important features of the methods are given.

AI.4.2 PCA for data overview

In PCA, data in the matrix **X** are transferred into a new co-ordinate system. This co-ordinate system is defined by the principal components. The direction in variable space occupied by the most varying data points will define the location of the first PC, and the second PC will coincide with the largest variation orthogonal (at right angle) to the first component. New PCs can be extracted in this way until only minor variation is left unexplained by the PC model. Each consecutive PC (a) consists of a score vector ($\mathbf{t}_a$) and a loading vector ($\mathbf{p}_a$'). The score vector contains one numerical value (a "score") for each observation that tells how far from the origin the observation lies in the direction of the loading vector. A PCA score plot shows the location of the observations relative to each other, and a loading plot displays how important the variables are for the different directions in a score plot. Observations close to each other in a score plot have similar properties and variables close to each other in a loading plot are also correlated.

Scores plotted in a scatter plot are useful for detecting strong outliers (Figure AI.6), clustering among observations (Figure AI.7), and time trends (Figure AI.8). Strong outliers may have a degrading impact on model quality and should normally be discarded. Grouped data may make it difficult to understand properties inside a cluster. Hence, a splitting of data into smaller groups according to the nature of the clustering is often done. Separate PCA models may then be fitted to better resolve the structure inside each cluster. In addition to a score plot, a plot of model residuals (DModX) is useful for uncovering weakly deviating observations that do not conform well with the dominant correlation structure (Figure AI.9).



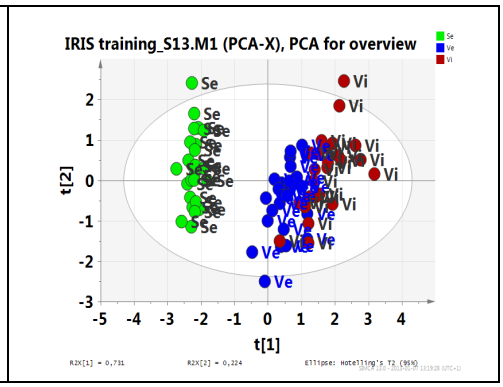
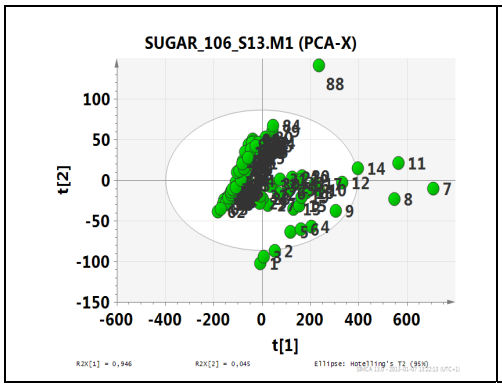


Figure AI.6: (left) PCA score plot of a process. Observation 39 is a strong outlier and has shifted PC1 relative to itself. Figure AI.7: (right) PCA score plot of a dataset with grouped observations. Here, it is recommended to split the dataset and carry out disjoint PCA modeling. This will help in resolving the clusters.



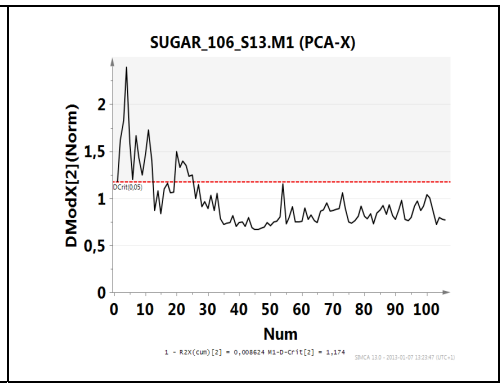


Figure AI.8: (left) PCA score plot of the SUGAR dataset. We can see that a marked process drift is manifested by the variables. Figure AI.9: (right) DModX chart for SUGAR PCA model, showing a typical short-term deviation (process upset) at the beginning of the sampling campaign. After a while, the process reaches a state of relative stability.

Furthermore, in process modeling, both scores and DModX may be plotted in control charts to better understand process behavior over time, and to see when a process tends to depart from a state of good operating conditions (Figure AI.10).

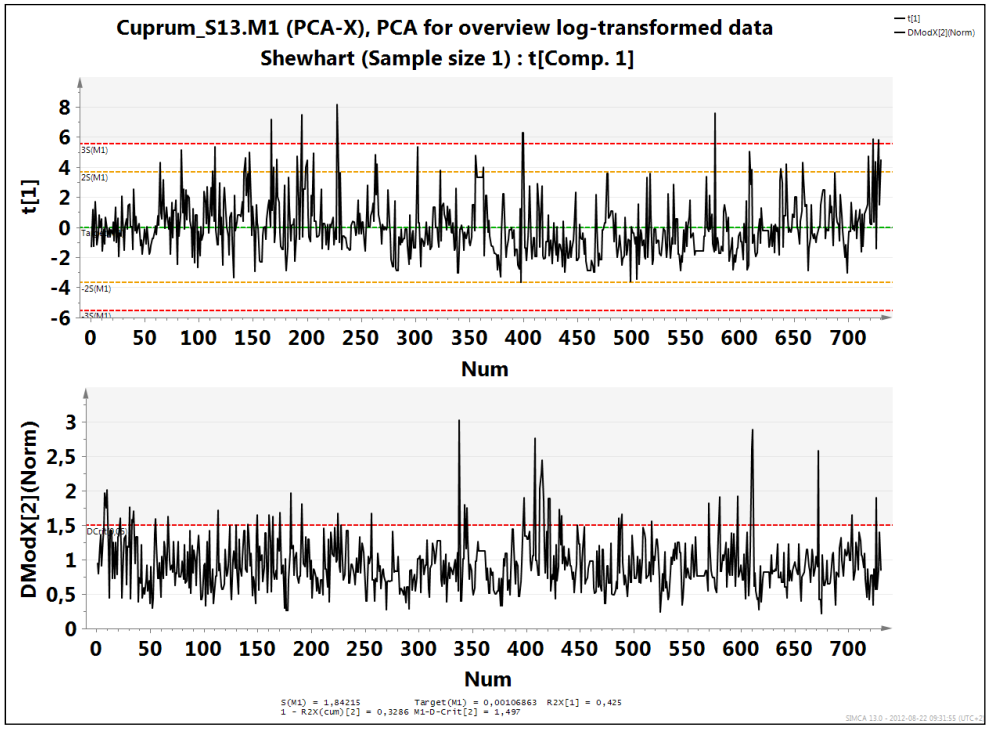


Figure AI.10: A Shewhart chart of the t_1 score of the PCA model for the CUPRUM dataset. The y-axis is the score value, and the x-axis represents sample number. The Shewhart plot of t_1 summarizes the systematic variation in the data. Tiled underneath is a plot of DModX, which enables an overview of the unsystematic process variation.

The PCA loadings may be plotted differently depending on the type of investigation. The scatter plot is practical in most situations and enables an understanding of all variables at the same time (Figure AI.11). Variables that are situated close together convey similar information about the observations, whereas variables that are farther apart carry dissimilar information. It is also possible to discover more or less uninfluential variables, because these are found close to the center of the co-ordinate system. The second kind of plot representation of loadings is exemplified in Figure AI.12. In the analysis of spectroscopic data, the plotting of each individual loading vector against the variable (wavelength) number is common practice, as this allows a direct comparison with spectral features.

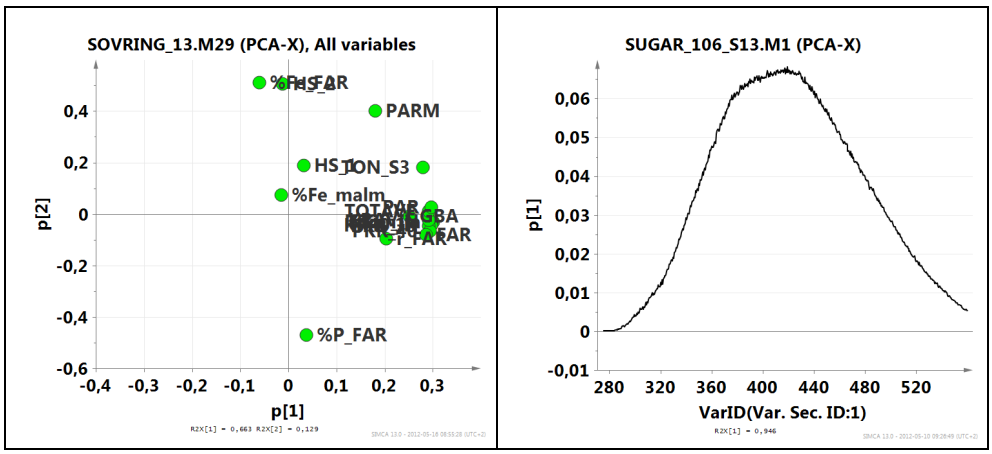


Figure AI.11: (left) PCA loading plot of the model for the SOVRING dataset. Figure AI.12: (right) PCA loading spectrum of the first component of the PCA model for the SUGAR dataset, which mimics the average fluorescence spectrum.

AI.4.3 PLS for quantitative modeling and predictions

In the case of PCA there is only one matrix (**X**) to consider. When making, e.g., a calibration model or a batch evolution model, one also has to consider a single response variable (**y**) or a matrix of responses (**Y**). The idea is that these response data should be modeled and predicted from the X-data. The X- and Y-matrices are decomposed in a similar way as in PCA. One fundamental difference is that while PCA captures maximum variance directions in the X-space, PLS finds directions in the X- and Y-spaces corresponding to maximum covariance.

One interpretation of PLS is that it forms "new X-variables", $\mathbf{t}_a$, as linear combinations of the old ones, and thereafter uses these new t's as predictors of Y. Only as many new t's (components) are formed as are predictively significant by cross-validation. For each component (a), the parameters, $\mathbf{t}_a$, $\mathbf{u}_a$, $\mathbf{w}^*_a$, $\mathbf{p}_a$, and $\mathbf{c}_a$, are determined by a PLS algorithm. For the interpretation of the PLS model, the scores, **t** and **u**, contain the information about the observations, and their similarities / dissimilarities with respect to the given problem and model. The weights **w*** and **c** give information about how the variables combine to form the quantitative relation between X and Y. Hence, these weights are essential for understanding which X-variables are important (numerically large w-values), which X-variables provide the same information (similar profiles of w_a-values), and for interpretation of the scores, t, etc.

A common PLS score plot is the t/u-plot. This plot is powerful in that it gives a visualization of the correlation between X and Y. When all observations are modeled well, a typical score plot may look like the one presented in Figure AI.13. Sometimes one or several observations may deviate in the relation between X and Y, and this is easily seen in a PLS score plot (Figure AI.14).

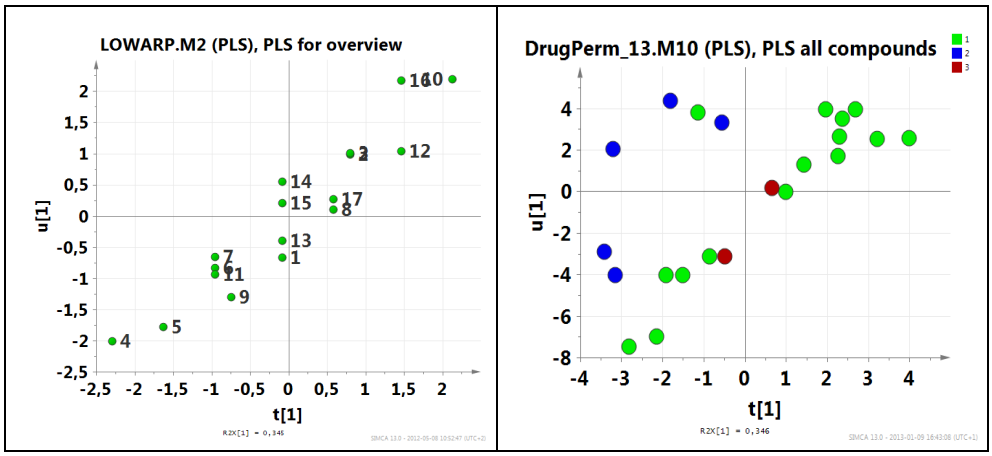


Figure AI.13: (left) PLS t_1/u_1 score plot of the model for the LOWARP dataset indicating a strong correlation between the factors and the responses. There are no outliers, i.e., off-diagonal points in the first dimension.
Figure AI.14: (right) PLS t_1/u_1 score plot of the model for the DrugPermeability QSAR dataset. The green dots denote compounds that are all passively transported drugs, and they dominate the model. Blue dots denote actively transported drugs, which deviate because their biological mechanism is different. Red dots show the positioning of compounds using paracellular transport facilities, observations that conform better with the major correlation structure.

In addition to detecting outliers in the relation between X and Y, the PLS t/u scores are useful for identifying departures from linearity between X and Y. Curvature between X and Y may appear in two ways. Figure AI.15 shows an example of what a t_1/u_1 score plot may look like when transformation of a response variable is necessary. The other main type of non-linearity between X and Y, involves a curvature of inverse arching (Figure AI.16). In this latter case, the solution is a transformation of X, e.g., the logarithmic transformation.

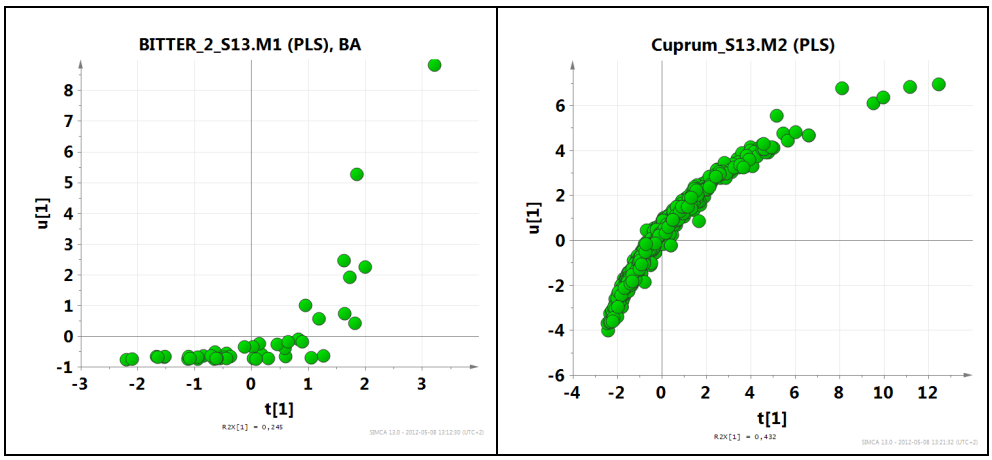


Figure AI.15: (left) QSAR example where a non-linear relationship is found between t_1 and u_1. By log-transforming the response variable (y), the inner relation in the PLS model becomes linear. Figure AI.16: (right) Process dataset where the inner relation is non-linear. By log-transforming the predictor variables (X), the inner relation in the PLS model becomes linear.

The part of the data that is not explained by the model, the residuals, is of diagnostic interest. Large residuals of Y (DModY) indicate that the model is poor, and a normal probability plot of the residuals of a single Y-variable, or a plot of DModY is useful for identifying outliers. In PLS we also get residuals for X, the part not used in the modeling of Y. These X-residuals (DModX) are useful for identifying outliers in the X-space, i.e., observations that do not harmonize with the model. Figures AI.17 and AI.18 show DModY and DModX plots pertaining to PLS models of the SOVRING and CUPRUM datasets, respectively.

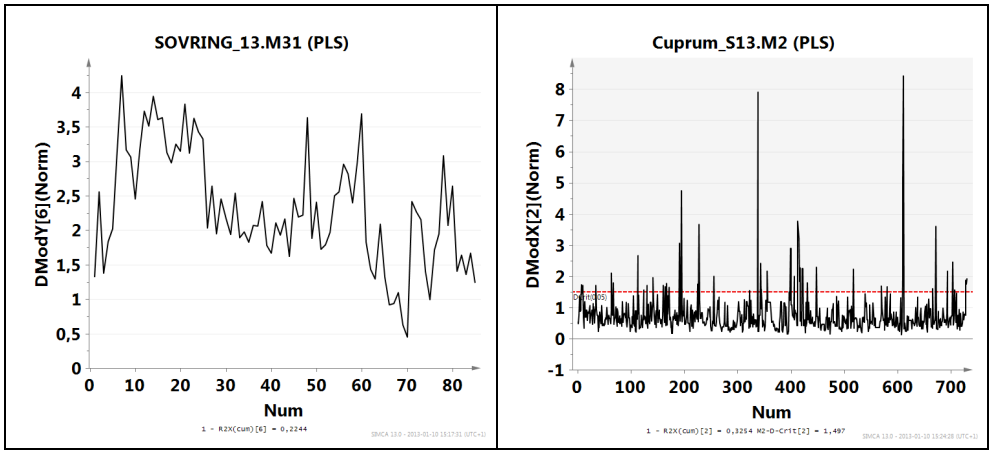


Figure AI.17: (left) DModY plot of a process where there are moderate outliers in the data. Rather many observations have Y-data differing from the majority of observations. Such a deviation is evidenced by a comparatively large DModY value. Figure AI.18: (right) DModX plot of a process where there are moderate outliers (spikes) in the data. The DModX values of several observations exceed the critical distance.

Finally, we address the interpretation of a PLS model, which is accomplished using either weights, coefficients, or VIP. Being a weighted sum of squares of the PLS weights, w*, and taking into account the amount of explained Y-variance of each dimension, the VIP parameter is the most condensed one for model interpretation. Its attractive property is the intrinsic parsimony; for a given model and problem there will always be only *one* VIP-vector, summarizing all components and Y-variables. Figure AI.19 shows a plot of VIP for the LOWARP dataset. We can see that two factors are more important than the others. However, in order to reveal the correlation structure between factors and responses at the same time, it is instructive to plot the PLS weights. Figure AI.20 shows the PLS weight plot of the LOWARP dataset.

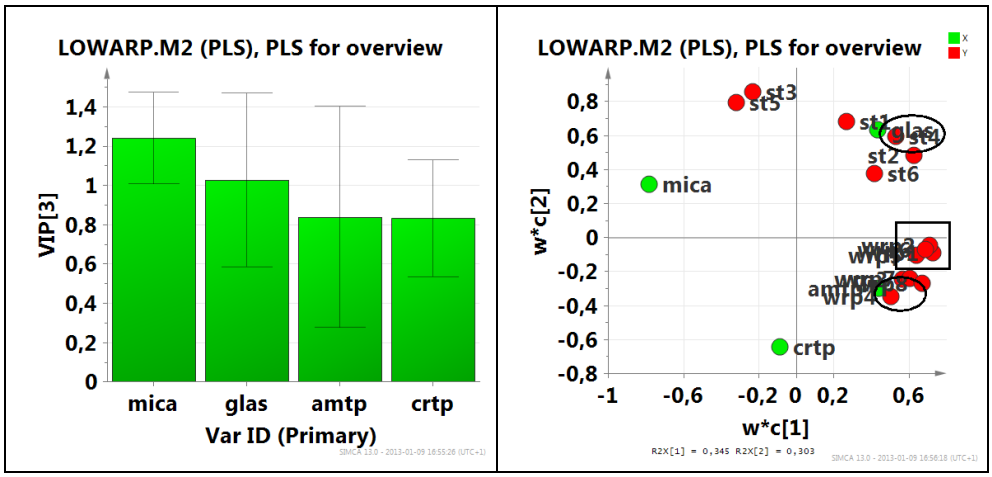


Figure AI.19: (left) PLS VIP plot of the model for the LOWARP dataset. The factors mica and glas are more important than crtp and amtp. Figure AI.20. (right) PLS weight plot of the model for the LOWARP dataset. For illustrative purposes, two uncorrelated responses are marked with ellipses and two correlated ones are marked with the box.

Alternatively, a PLS model may be interpreted by means of the regression coefficients. The LOWARP dataset comprises 14 responses, and it is therefore necessary to inspect 14 sets of coefficients. For reasons of brevity only four such coefficient sets are displayed (Figures AI.21 – AI.24). The four sets of coefficients which are plotted relate to the four responses highlighted in the weight plot (Figure AI.20), viz., wrp1, wrp2, wrp4, and st4. As expected, the two correlated responses (wrp1 and wrp2) demonstrate very similar coefficient profiles (Figures AI.21 and AI.22), whereas the corresponding profiles for the two uncorrelated responses (wrp4 and st4) differ greatly (Figures AI.23 and AI.24).

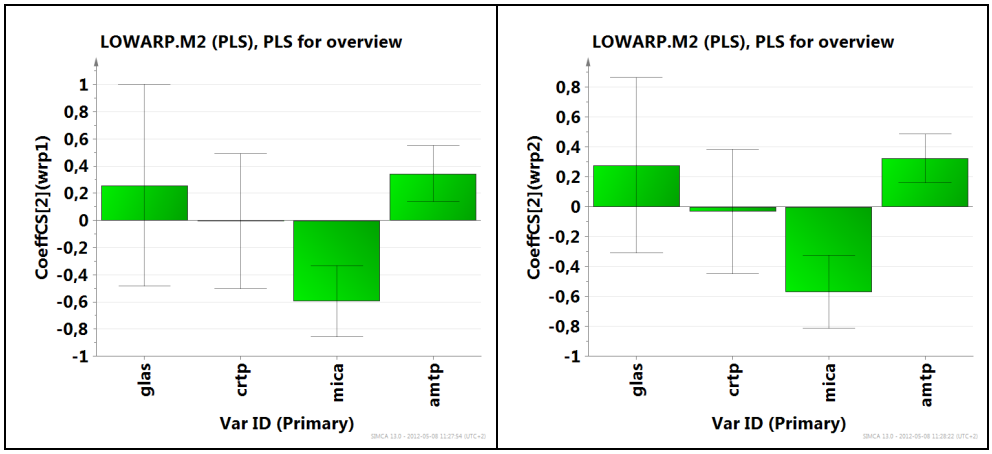


Figure AI.21: (left) Regression coefficients of scaled and centered variables for the response wrp1. Figure AI.22: (right) Regression coefficients of scaled and centered variables for the response wrp2.

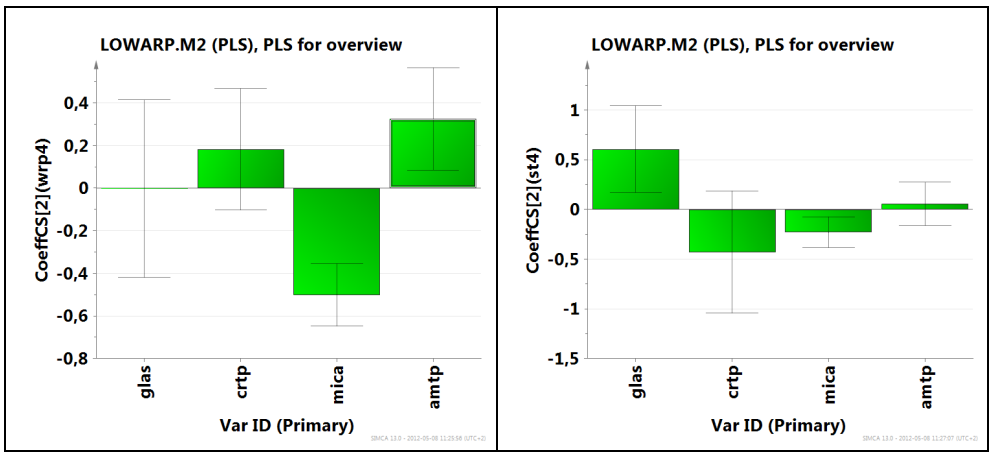


Figure AI.23: (left) Regression coefficients of scaled and centered variables for the response wrp4. Figure AI.24: (right) Regression coefficients of scaled and centered variables for the response st4.

AI.4.4 OPLS for enhanced interpretation

OPLS is an extension of PLS that makes model interpretation easier. Figure AI.25 sheds some light on the difference between PLS and OPLS. PLS operates by dividing the variability – the sum of squares – in the X-block in two parts, i.e., the systematic and residual parts. OPLS further splits the systematic variability, R2X, in two parts, the part that is correlated (predictive) to Y and the part that is uncorrelated (orthogonal) to Y.

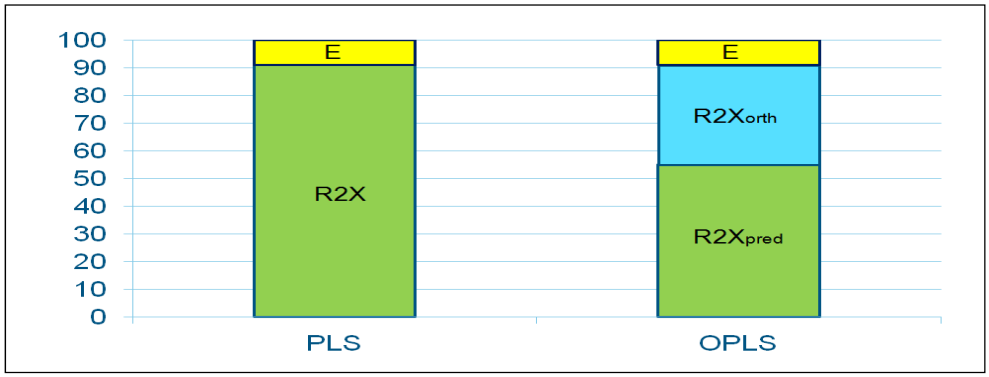


Figure AI.25: PLS (left) divides the variability in the X-matrix in two parts, the systematic variability and the residual variability. OPLS (right) further splits the systematic variability, R2X, in two parts, the part that is correlated (predictive) to Y and the part that is uncorrelated (orthogonal) to Y.

Frequently, in the single-Y case, the immediate effect of using OPLS is a rotation of the corresponding PLS score space such that the predictive variation is enriched into the first dimension of the OPLS model. Figure AI.26 shows the first score plot of the PLS and OPLS models for the BeerNIR dataset discussed in Chapter 5. Two dominant directions, corresponding to color and alcohol content, are found. The predictive component of the OPLS model is well aligned with the alcohol content, whereas the corresponding structure is not accounted for in the first component of the PLS model. This shows the inclination of PLS to pick up strong systematic structures in X in the first component that are not necessarily fully relevant for the Y-variable(s) under study. The added advantage of separating the predictive and orthogonal variations into separate components is that the analyst gets a better understanding of what is captured by the various components.

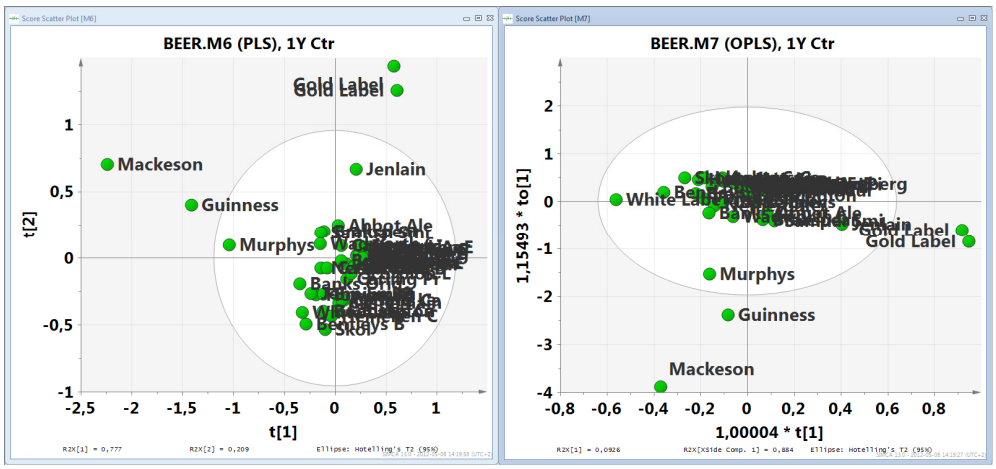


Figure AI.26: Score plots of the PLS (left) and OPLS (right) models. The two models have picked up the two dominant structures in the spectral data, color and alcohol content, but in the OPLS model the predictively relevant information is concentrated into the first component.

All projection methods working in accordance with a least squares approach are sensitive to abnormalities in data. It is important to realize that detection and investigation of outliers represents an important part of multivariate data analysis and semi-empirical modeling. As discussed in Chapters 3 and 4, in both PLS and PCA, abnormal samples can be detected and analyzed by looking at plots of scores, residuals and contributions. Outlier detection in OPLS is no different because exactly the same principles apply.

AI.4.5 O2PLS for data integration

Another modification to the standard PLS method is denoted O2PLS. The primary objective of O2PLS is integration of data in the two-block (X/Y) context and understanding of which information overlaps between the two data tables and which information is unique to a specific data table (X or Y). O2PLS computes (i) components expressing the joint X/Y information overlap, (ii) components expressing what is unique to X, and (iii) components expressing what is unique to Y.

Moreover, the switch in operational focus between OPLS (regression between two blocks of data) and O2PLS (integration of two blocks of data) is accompanied by a transition in the preferred terminology. Whereas with OPLS it is practical to distinguish between *predictive* and *orthogonal* sources of variability, when using O2PLS it becomes natural to consider *joint* and *unique* information structures. The joint variation represents the information overlap between X and Y, i.e., the systematic structure that is present in both X and Y. Any unique variation corresponds to information beyond the overlap and which is exclusively found in either the X- or the Y-block. An overview of how O2PLS divides X- and Y-variabilities is displayed in Figure AI.27.

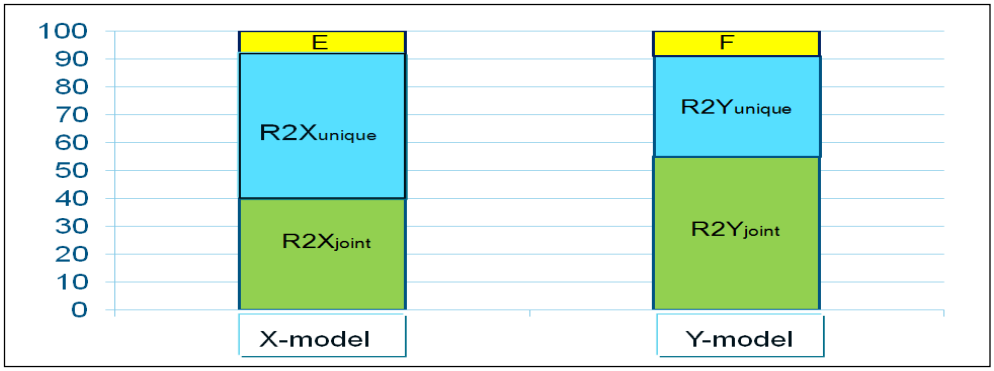


Figure AI.27: O2PLS finds the joint variation in X ($R2X_{joint}$) and Y ($R2Y_{joint}$), which represents the information overlap between X and Y. Furthermore, O2PLS finds the unique variation in X ($R2X_{unique}$) and the unique variation in Y ($R2Y_{unique}$), which correspond to information that is exclusively found in either X or Y. Finally, the X- and Y-variabilities that O2PLS cannot cope with go to the E and F residual structures.

As discussed in Chapter 6, the various compartments of the O2PLS model, and the corresponding residuals, are interpretable using the usual scores, loadings, DModX and DModY parameters.

AI.5 Model validation and use of model

AI.5.1 Main steps

Before a model is "released" for use, it is mandatory to carry out model validation. First of all, it is important to make clear what we mean by a *valid* model. We mean that it predicts much better than chance. In addition, it should have model coefficients with little bias, i.e., they should have the correct sign and be large for important variables and small for unimportant variables. Finally, it should be consistent with fundamental chemical/biological/engineering knowledge.

The main steps in *predictive* model validation are:

- Internal validation (cross-validation)
- Response permutation testing
- External validation
- Predictions of new observations

AI.5.2 Why is predictive validation necessary at all?

In any modeling, including multivariate modeling, it is easy to manipulate data such that an apparently good model can be formulated. The most drastic step here is removal of observations and variables that "do not fit" according to some subjective criterion. Furthermore, variables might be unduly transformed, and model complexity (i.e., number of latent variables) might be driven beyond pertinent limits. Such an inappropriate model often arises when one is merely interested in the fit of the model to the underlying data, and neglects its performance with new samples. The problem with this kind of model is that it is not representative for other, additional samples. *Predictive validation* is the only way to reliably assess model adequacy for new observations.

In this context, it is of crucial importance to realize the difference between a model's fit (R^2) and prediction ability (Q^2). The fit tells how well we are able to mathematically reproduce the (response) data of the training (calibration) set. A quantitative measure of the goodness of fit is given by the parameter R^2 (see definitions of R^2X and R^2Y in Chapters 3 and 4, respectively). The problem with the goodness of fit is that with sufficiently many free parameters in the model, R^2 can be made arbitrarily close to the optimal value of one (1.0). Fortunately, the prediction ability is not as easy to manipulate. It measures how accurately we can predict the data of new observations not previously used in the model training. The predictive power of a model may be estimated by the goodness of prediction parameter Q^2 (see definitions in Chapters 3 and 4).

An important aspect in predictive validation of a model concerns the composition of the training (calibration) set. In classification, calibration, and process modeling, it is of utmost importance to use training set observations that are representative for the group of samples under investigation. In reality it may be difficult to obtain representative samples in a short time, e.g., due to seasonal variations of raw material, or expensive and time-consuming testing at the quality control laboratory. Nevertheless, it is important that as many informative and representative samples are gathered as is practically possible. One problem with too few training set observations, or observations that are not representative, is that the confidence intervals in a PCA/PLS/OPLS/O2PLS model will tend to be too conservative. Hence, it might be that observations which really belong to the class fall outside the model limits. If new observations are detected that have the "right" properties, these should be included in the model. This updating has to be continued until the class model becomes stable and correctly assesses the class membership of new observations.

As should be apparent from the discussion above, predictive validation and training set composition are intimately linked. The problem of how to compose a decent training set has been discussed in several chapters in this book (see e.g. Chapters 7, 8, and 18). In the following sections we consider some tools for predictive validation. The most demanding way to validate a model is by external validation, which consists of making predictions for an independent set of data not used in the model calibration. Ideally, such a prediction set should be selected according to a statistical experimental design. However, external validation might not always be tractable, e.g., in QSAR because of the resources needed to make a test set of new compounds. Hence, alternatives for predictive validation are of interest. The additional model validation alternatives discussed here are internal validation (cross-validation), and cross-validation together with permutation testing.

AI.5.3 Internal validation (cross-validation)

In Sections 3.5.3 and 4.7.3 the use of cross-validation for significance testing of each consecutive PCA or PLS component was described. Not only does cross-validation give an estimate of the significance of a latent variable, but it also provides an estimate of the overall predictive power of a model.

Basically, cross-validation (CV) is performed by dividing the data into a number of groups and then developing a number of parallel models from the reduced data with one of the groups deleted. It should be noted that making the number of CV-groups equal to N, i.e., the so called leave-one-out (LOO) approach, is a dubious practice [Shao, 1993], because LOO gives too small a perturbation to the data, so that Q^2 approaches the properties of R^2. For designed data, however, LOO might be a viable option. In general, between five and ten CV-groups works well.

After developing a model, the deleted data are used as a test set, and differences between actual and predicted Y-values are calculated for the test set. The sum of squares of these differences is computed and collected from all the parallel models to form PRESS (predictive residual sum of squares), which is a measure of the predictive ability of the model. Usually, PRESS is re-expressed as Q^2 (the "cross-validated R^2") which is (1-PRESS/SS) where SS is the sum of squares of Y, corrected for the mean.

The companion parameters R^2 and Q^2 are informative for surveying the model performance and evolution. Our illustration to this point relates to the PLS model of the LOWARP dataset. Figure AI.28 shows the evolution of R^2Y and Q^2Y for the respective model dimensions of the PLS model. In this case it seems relevant to use two components, because the predictive ability does not increase to any appreciable extent beyond the second dimension.

By investigating the R^2Y's and Q^2Y's of the individual responses it is possible to see which responses are well modeled, and which are not (Figure AI.29). Figure AI.29 presents the situation after two PLS components. Apparently, there are strong, quantitative relationships between the factors and most of the responses.

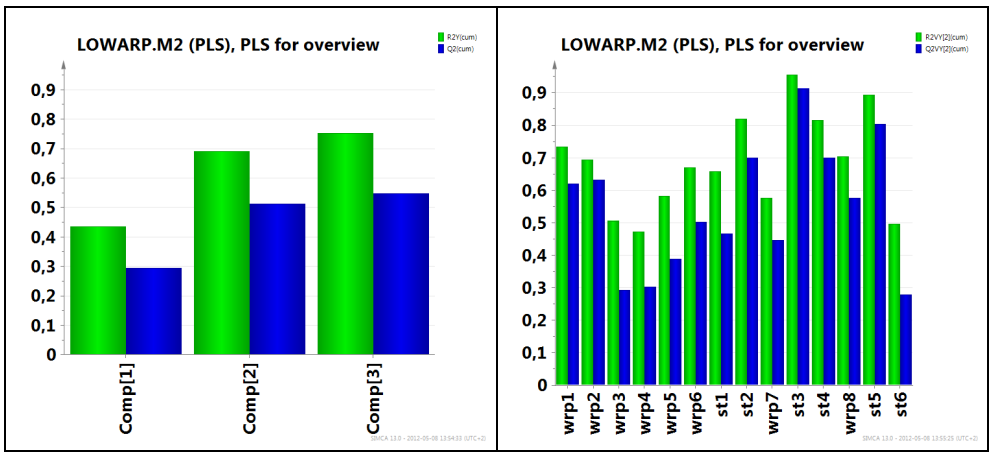


Figure AI.28: (left) Model overview of the PLS model for the LOWARP dataset. The plot shows how the overall R^2Y and Q^2Y-statistics change as a function of increasing model complexity. Here, two components appear appropriate, as Q^2Y increases only marginally with the introduction of the third component. Figure AI.29: (right) The overall R^2Y and Q^2Y-values may be decomposed into R^2's and Q^2's relating to the individual responses. The situation after two model components is depicted. Most variables are rather well modeled after two PLS components, but there are some exceptions. Particularly, one should analyze the outcome for variables like wrp4 and st6 more stringently, in order to understand what is causing their poor modeling.

We feel that it is important to point out some cases where cross-validation is known to face problems. This occurs with (i) designed data, (ii) grouped and/or duplicated data, and (iii) time-dependent process data. The first problem, related to DOE-data, is that when parts of the data are removed in CV, the underlying design often collapses and there is no longer any preferred projection direction. This may lead to unrealistically low estimates of the predictive power. The second problem, linked to grouped data, is that although many samples are omitted in each CV-cycle, strong groupings may still persist and thereby cause the estimate of the predictive power to be higher than it should be. The third problem, related to auto-correlated process data, is that the predictive power may not be reliably estimated, because time points (observations) adjacent to those eliminated in a CV-round, carry information very similar to the eliminated data points. Hence, CV with auto-correlated process data may lead to overly optimistic Q^2-values. One alternative here might be to employ response permutation, or to sort the data according to Y to break up the auto-correlation among adjacent samples.

AI.5.4 Response permutation and cross-validation

In this section, we will describe a diagnostic tool which we have not addressed before, namely the *response permutation* option in SIMCA.

One limitation of cross-validation is that it assesses only the predictive power, but provides no statement of the statistical significance of the estimated predicted power. In order to obtain an estimate of the significance of a Q^2Y-value, one may in PLS develop a number of parallel models based on fit to randomly re-ordered Y-data, and evaluate the real Q^2Y in the light of a distribution of Q^2Y-values of re-ordered response data. This option is called *response permutation* in SIMCA. A good description of permutation testing can be found in [van der Voet, 1994].

Validate works as follows: For the training set, the X-data are left intact, while the Y-data are permuted to appear in a different order. This means that the Y-data remain numerically the same, but their positions are shifted by random shuffling. A PLS model is then fitted to the permuted Y-data, and by using cross-validation both R^2Y- and Q^2Y-values are computed for the derived model. These "permuted" values may then be compared with the estimates of R^2Y and Q^2Y of the "real" model to get a first indication of the significance of the latter values.

In the next round, a second PLS model is fitted to another permuted version of the Y-data, and new estimates of "permuted" R^2Y- and Q^2Y-values are thus formed. By repeating this permutation procedure a number of times, say, between 25 and 100, and by establishing an

equivalent number of parallel PLS models, it is possible to achieve reference distributions of R^2Y and Q^2Y based on random data. Such reference distributions are useful for appraising the statistical significance of the R^2Y- and Q^2Y-parameters of the parent PLS model.

Furthermore, because the numerical values of the "permuted" versions of R^2Y and Q^2Y depend, at least partly, on the extent of perturbation inflicted by the permutation procedure, it is advisable to keep track of the correlation coefficient between original and permuted Y-variables. Should an original Y-variable be only mildly perturbed by permutation, there will by necessity be a high correlation coefficient between the permuted and the original variable. By jointly assessing such correlation coefficients and "permuted" R^2Y/Q^2Y-numbers, it is possible to understand and explain the existence of occasionally high R^2Y- and Q^2Y-values for permuted Y-data.

In many cases of weak PLS models, however, conspicuously high R^2Y- and Q^2Y-estimates for permuted response variables cannot be explained by an incidentally high correlation between permuted and original Y-data. This then constitutes an indication of an inappropriate and perhaps even overfitted model.

The *response permutation* plot in SIMCA is a recent and informative way of summarizing results of response permutation testing. In Figure AI.30, validate plots are given for two responses of the LOWARP dataset, the well modeled response st3 and the poorly modeled response st6. This kind of plot is constructed by letting the Y-axis represent the R^2Y/Q^2Y-values of all PLS models, including the "real" one, and by assigning the X-axis to the correlation coefficients between permuted and original response variables. Observe that the points of R^2Y and Q^2Y for the original model are always found in the right-hand part of each plot at correlation 1.0 (because 1.0 is the correlation coefficient obtained when correlating a variable with itself).

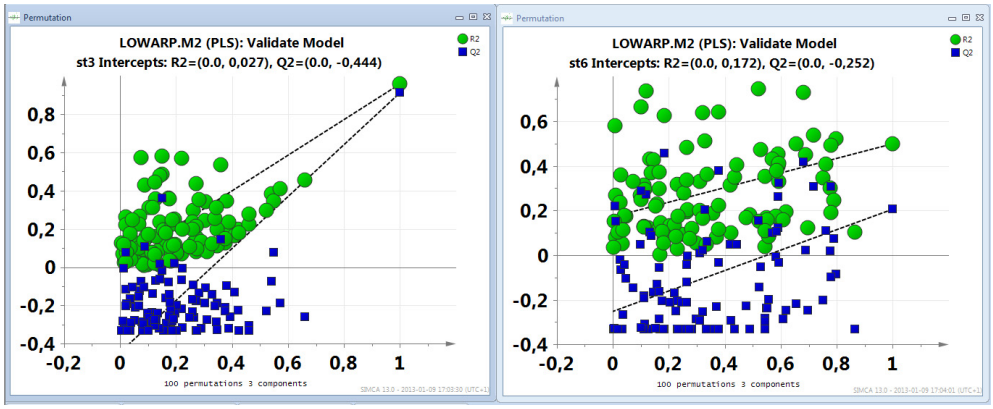


Figure AI.30: Validate plots of responses st3 (left) and st6 (right) of LOWARP dataset. The Y-axis represents R^2Y (triangles) and Q^2Y (squares) for every model, and the X-axis designates the correlation coefficient between original and permuted response data. See text for explanation of plot interpretation.

As shown by Figure AI.30, R^2Y and Q^2Y of the original st3 model are always substantially higher than the corresponding "permuted" values. This indicates model validity with regard to st3. However, there is a considerable spread among the permuted R^2Y's and Q^2Y's of st6, indicating that the PLS model obtained is weaker for this response. It would be desirable to somehow summarize the results of permutation testing in a quantitative manner. One way of doing this is to conduct conventional regression analysis in the two sets of points, i.e., one regression line is fitted among the R^2Y-points (green circles) and another line among the Q^2Y-points (blue squares). The intercepts of the resulting regression lines are interpretable as measures of "background" R^2Y and Q^2Y obtained by fit to random data. Experience shows that the R^2Y-intercept should not exceed 0.3-0.4 and that the Q^2Y-intercept should not exceed 0.05. Intercepts below these limits indicate valid models. Returning to the validate plots of Figure AI.30, we may now see that the PLS model is valid for st3. The results for st6 are also good according to the reference limits for intercepts, but the modeling of st6 is weaker than for st3.

The response permutation testing procedure may also be used for tentative assessment of PLS model complexity. Consider the validate plots shown in Figures AI.31 and AI.32, which show validate plots relating to different complexities of a PLS model of the reduced SUGAR dataset (compare with discussion in Section 8.4.4). Depending on how cross-validation results are interpreted either two or five PLS components are conceivable in the modeling of the ash content. For reasons that will become apparent shortly, the modeling results shown in Section 8.4.4 refer to two model dimensions. The results of the permutation testing (after 100 permutations) of the two-component model are plotted in Figure AI.31. For the corresponding five-component model, permutation testing results are plotted in Figure AI.32. Clearly, the two-component model is the preferable and more valid one. The five-component model is highly overfitted. External validation also indicated the appropriateness of two latent variables.

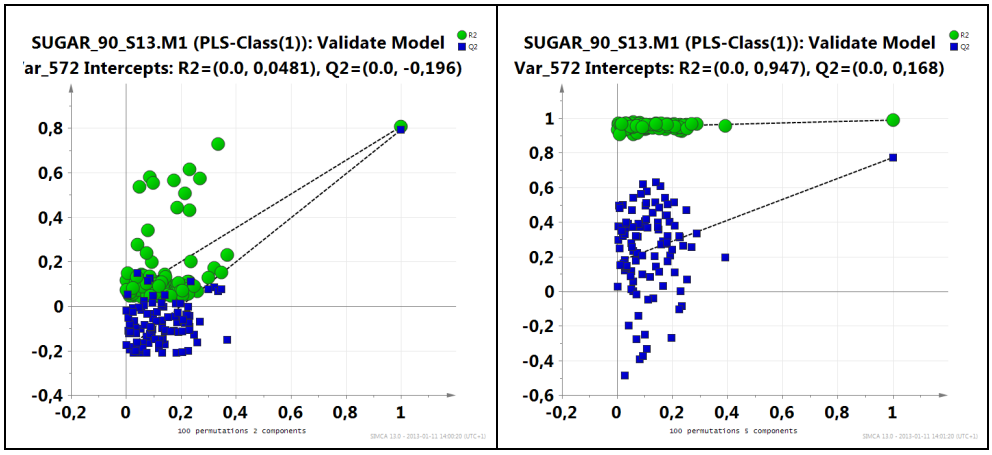


Figure AI.31: (left) Validate plot (after 100 permutations) of the two-dimensional PLS model of the reduced SUGAR dataset. Figure AI.32: (right) Validate plot (after 100 permutations) of the five-dimensional PLS model of the reduced SUGAR dataset.

Thus, in summary, the combination of cross-validation and response permutation testing represents an interesting alternative for model validation. Cross-validation assesses the predictive power of a model, and permutation testing assesses the statistical significance of the estimated predictive power.

AI.5.5 External validation

Predictive validation by means of cross-validation and response permutation testing in many ways provides a reasonable first approximation of the predictive ability of a PLS model. However, a more demanding and rigorous way of testing predictive performance consists of computing predictions for an independent set of test observations. In this respect, an independent test set means a series of observations, neither used in the model building process nor presented to the analyst until the final version of the current model has been postulated.

In order to accomplish a proper external validation it is essential to understand the nature of the data. The training set and the test set must be representative for all those variations in X-data that may occur during normal, productive operation of the system or process under study. A convincing presentation regarding the crucial importance of a properly selected training set in multivariate modeling of spectroscopic data of chemically modified celluloses, was given in the work of Svensson and co-authors [Svensson, *et al.*, 1997]. In addition, in PLS modeling, it is essential that the training- and test sets span the range of the response variables fully. The best way to obtain good training- and test sets is through the use of multivariate design. Also, it should be noted that a couple of iterations of model updating are often necessary to achieve the goals of representativity and information richness.

AI.5.6 Predictions of new observations (Use of model)

Once the PLS model has been estimated from the training set of X and Y, the model can be used to predict Y from X for new observations. Similarly, a PCA model can be utilized to predict class membership of new observations. For many crucial applications, especially expensive on-line applications and pharmaceutical applications, the original method (giving the Y-values) is run in parallel with the new method (giving the X-values) to establish that the calibration is working satisfactorily during regular use. This validation can often include the change of personnel, equipment, batches of raw materials, and other factors including seasonal factors that might affect the calibrations. Within the agricultural area it is more or less mandatory to update the model for each season. After the model has been validated in this extensive way, which can take months, it will be used. While using the model it is still recommended that one sample out of 10, 100, 1000, 10000 is measured with the original method.

Furthermore, all the available PCA and PLS parameters and diagnostic tools should be used in order to understand whether the new observations conform to a certain model or not. Should a new observation for some reason be different from the training samples, it is of course possible, but very risky, to make predictions. The use of model residuals on the X-side (DModX) makes is possible to classify a new observation as similar or dissimilar to the training set observations, and thereby to get an insight into model relevance. Moreover, in PLS, it is also possible to compute uncertainty measures (standard error, confidence intervals) of predicted response data.

It is very important in at-, in-, and on-line analysis to check the relevance of predicted values. If it is possible to detect deviations from normal behavior or unrepresentative samples of the process while it is still running, there is a possibility for an operator, or an automatic feed-back controller, to adjust the process back to target. Such deviations from normal performance may arise from varying quality of raw material, and deviations in e.g. temperatures and pressures.

AI.6 Questions for Appendix I

1. What are the main steps of multivariate data analysis?
2. Why is MVDA often of an iterative nature?
3. Which PCA/PLS/OPLS/O2PLS parameters are useful for model interpretation?
4. Which tools are available for model validation?
5. Which feature is assessed with response permutation?

AI.7 Summary and discussion

AI.7.1 MVDA is an iterative procedure

In this appendix we have emphasized the iterative nature of MVDA. We have seen that PCA, PLS, OPLS and O2PLS are powerful data analytical techniques for multivariate data analysis. However, in many cases these methods are not used directly on raw data just as they appear in a data base. On the contrary, raw data are often scaled, centered, transformed, corrected and compressed prior to model building. Therefore, the step of data pre-processing is intimately linked to the data analysis step. Moreover, PCA, PLS, OPLS and O2PLS give rise to many model parameters and diagnostic tools, which may be plotted during model review and interpretation.

Appendix II: Statistics

Introduction

This chapter contains brief background information and formulas for SIMCA. For more extensive reading, use the reference material.

The topics covered are:

PCA - Principal Components modeling

PLS - Partial Least Squares Projection to Latent Structures modeling

OPLS/O2PLS - Orthogonal PLS modeling

Cluster Analysis (CA), dendrograms, Hierarchical CA (HCA), PLS including Hierarchical Cluster Analysis - HCA and PLS-Trees

Vectors available in SIMCA

Formulas and descriptions

Transform page criteria

Scaling

Cross validation

PLS Time Series Analysis

CV-ANOVA

Fisher's Exact test

Control Chart statistics

S-plot

Fit methods background

A dataset is composed of N rows and K columns. The N rows are here called observations and the K columns are called variables.

Geometrically we can represent the observations as points in a multidimensional space where the variables define the axes. The lengths of the axes are determined by the scaling of the variables.

To calculate a model that approximates the dataset the following fit methods are available in SIMCA:

PC modeling.

PLS modeling.

OPLS and O2PLS modeling.

All other fit methods, available by clicking **Change Model Type** use one of the fit methods above, see the subsection in the **Home** chapter.

PCA - Principal Components modeling

Principal Component Analysis is the technique for finding a transformation that transforms an original set of correlated variables to a new set of uncorrelated variables, called principal components. The components are obtained in order of decreasing importance, the aim being to reduce the dimensionally of the data.

The analysis can also be seen as an attempt to uncover approximate linear dependencies among variables.

PC modeling shows the correlation structure of your data matrix X, approximating it by a matrix product of lower dimension (**TP'**), called the principal components plus a matrix of residuals (**E**).

X = Xbar + TP' + E

where

Xbar contains X average.

T is a matrix of scores that summarizes the X-variables.

P is a matrix of loadings showing the influence of the variables.

E is a matrix of residuals; the deviations between the original values and the projections.

This geometrically corresponds to fitting a line, plane or hyper plane to the data in the multidimensional space with the variables as axes. The scaling of the variables specifies the length of the axes of this space. For more about scaling see the **Scaling** section later in this chapter.

SIMCA iteratively computes one principal component at a time, comprising a score vector $\mathbf{t}_a$ and a loading vector $\mathbf{p}_a$.

Number of model dimensions (A)

To get an overview of the dataset, a few (2 or 3) principal components are often sufficient.

However, if the PC model is used for modeling or for other predictions (e.g. principal properties), cross validation (CV) should be used for testing the significance of the principal components. In this way, the "significant" number of PC components, A, is obtained, which is essential in modeling.

It is also possible to use the eigenvalue limit (EV) for testing component significance by clearing the **Use cross validation when fitting** check box in **Model Options**, tab **Model**.

Eigenvalue Limit (EV): A component is considered significant if its normalized eigenvalue is larger than 2.

Cross validation (CV) is described in the **Cross validation** section later in this chapter.

Reference PCA

Jackson, J.E. (1991), A User's Guide to Principal Components, John Wiley, New York. ISBN 0-471-62267-2.

PLS - Partial Least Squares Projection to Latent Structures modeling

When fitting a Partial Least Squares Projection to Latent Structures model, PLS finds the **linear (or polynomial) relationship between** a matrix Y (dependent variables) and a matrix X (predictor variables) expressed as:

Y = f(X) + E

The matrix X refers to the predictor variables and their squared and/or cross terms if these have been added. Active X variables or expansions participating in the model are sometimes referred to as **terms**. The function f(X) is usually a polynomial, possibly in the transformed variables.

***Note**: Cross terms should be added to the linear terms in X **only** when the data supports such terms, for instance after using a design to generate X.*

PLS is most easily understood geometrically, where we see the matrices X and Y as **N** points in two spaces, the X-space with **K** axes, and the Y-space with **M** axes, **K** and **M** being the number of columns in X and Y.

PLS modeling consists of simultaneous projections of both the X and Y spaces on low dimensional hyper planes. The coordinates of the points on these hyper planes constitute the elements of the matrices T and U. The analysis has the following **objectives:**

To well approximate the X and Y spaces

To maximize the correlation between X and Y

The PLS model accomplishing these objectives can be expressed as:

X = Xbar + TP' + E

Y = Ybar + UP' + F

U = T + H (the inner relation)

where

Xbar contains X average.

Ybar contains Y average.

Note that the coefficients of the inner relation are 1.

In the PLS algorithm there are additional loadings, W, called weights. These express the correlation between U and X and are used to calculate T.

This modeling geometrically corresponds to fitting a line, plane or hyper plane to both the X and Y data represented as points in a multidimensional space, with the objective of well approximating the original data tables X and Y, and **maximizing the covariance** between the observation positions on the hyper planes.

SIMCA will iteratively compute one PLS component at a time, that is: one vector each of X-scores t, Y-scores u, weights w and c, and loadings p.

The PLS components are calculated in descending order of importance.

Vector	Description
T	Matrix of scores that summarizes the X variables.
W	Matrix of weights expressing the correlation between X and U (Y).
U	Matrix of scores that summarizes the Y variables.
C	Matrix of weights expressing the correlation between Y and T (X).
E, F, H	Matrices of residuals.

Number of model dimensions (A)

The criterion used to determine the model dimensionality, (number of significant PLS components), is cross validation (CV). With CV, observations are kept out of the model development, then the response values (Y) for the kept out observations are predicted by the model, and compared with the actual values.

This procedure is repeated several times until every observation has been kept out once and only once.

Reference PLS

Wold, S., Sjöström, M., and Eriksson, L., (2001b), PLS-Regression: A Basic Tool of Chemometrics, Journal of Chemometrics, 58, 109-130.

OPLS/O2PLS - Orthogonal PLS modeling

OPLS - Orthogonal PLS modeling

A simple way to understand OPLS is to consider how PLS and OPLS differ in their handling of the variance of the X-matrix. PLS divides the variability in X in two parts (figure below), i.e., the systematic and residual parts. The systematic part is the sum of the variability in X that is correlated (predictive) to Y and the variability in X that is uncorrelated (orthogonal) to Y. Thus, whereas PLS

divides the sum of squares of X in two parts, OPLS divides it in three parts. This yields good predictions and improved interpretability.

It should be noted that in the single-Y case, by theory, the OPLS model can only have one predictive component [Trygg and Wold, 2002]. Should a single-Y OPLS model comprise more than one component, all components beyond the first one reflect orthogonal variation. However, with multiple Y-variables there can be more than one predictive OPLS component.

For the single-y case only OPLS is available in SIMCA. The results from OPLS/O2PLS in SIMCA-P+ 12 are identical to the results using OPLS in SIMCA 13.

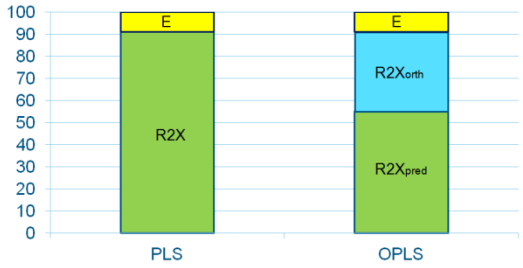


Figure: PLS (left) divides the variability in the X-matrix in two parts, the systematic variability and the residual variability. OPLS (right) further splits the systematic variability, R2X, in two parts, the part that is correlated (predictive) to Y and the part that is uncorrelated (orthogonal) to Y.

In the case of a single response variable, we can write the X-part of the OPLS model as

$X = 1x' + tp' + T_oP_o' + E$

and the OPLS model prediction of y as

$y = y' + tq' + f$

The multi-Y OPLS model can be expressed as follows:

$X = 1x' + TP' + T_oP_o' + E$

and the OPLS model prediction of Y as

$Y = 1y' + TQ' + F$

Where the matrix products TP' and TQ' hold the joint X/Y information overlap. The number of score vectors (in T and U) and loading vectors (in P' and Q') is determined using cross-validation.

Thus, in comparison with the single-y OPLS model, the main difference lies in the fact that there can be more than one predictive component in the multi-y OPLS model. The number of predictive components is regulated by the number of latent variables in the information overlap between X and Y, which in turn often is linked to the rank of Y.

O2PLS - Orthogonal PLS modeling

The O2PLS model can be written as follows:

$X = 1x' + TP' + ToPo' + E$

$Y = 1y' + UQ' + UoQo' + F$

Where the matrix products TP' and UQ' hold the joint X/Y information overlap. The number of score vectors (in T and U) and loading vectors (in P' and Q') is determined using cross-validation. In the example in the **Model Window for OPLS and O2PLS models** subsection in the **View** chapter, there are seven components of this type (predictive components).

The number of components in the respective set of components is determined using cross validation.

For any part of the OPLS/O2PLS model, the percentages explained and predicted variances can be obtained from plots and lists in the software.

For more about the differences between OPLS/O2PLS and PLS, see the **Conventional PLS compared with OPLS and O2PLS** subsection.

Conventional PLS compared with OPLS and O2PLS

PLS

Conventional PLS applies to the two-block (X/Y) regression problem. It uses X to construct a model of Y, where the objective is to predict the latter from the former for new samples in the predictionset. In that sense, PLS is unidirectional, i.e., X → Y, but not vice versa.

When X is composed of e.g. spectroscopic data, process readings or measurements from bio-analytical platforms, there is a risk that systematic variation may reside in X which is not linearly correlated with Y. Such variability in X is usually called Orthogonal in X [1]. Although the Orthogonal in X variation can be handled by a PLS model, it often makes model interpretation more difficult [2].

OPLS

The OPLS method is a recent modification of the PLS method [1-3], which is designed to handle variation in X that is orthogonal to Y. OPLS separates the systematic variation in X into two parts, one that is linearly related (and therefore predictive) to Y and one that is orthogonal to Y. The predictive variation of Y in X is modeled by the predictive components. The variation in X which is orthogonal to Y is modeled by the orthogonal components. This partitioning of the X-data provides improved model transparency and interpretability, and gives very similar predictive power. For OPLS with one Y the predictive model with the same number of components is identical to the same PLS model. Similarly to PLS, OPLS is a unidirectional method, where the scope is the relation X → Y.

O2PLS

O2PLS is a generalization of OPLS [4,5]. In contrast to PLS and OPLS, O2PLS is bidirectional, i.e. X ↔ Y. Additionally, with O2PLS it is possible to partition the systematic variability in X and Y into three parts, (i) the X/Y joint predictive variation, (ii) the variation in X orthogonal to Y (X-unique Variation), and (iii) the X-unrelated variation in Y (Y-unique Variation). The variation in X orthogonal to Y can be further divided into two parts; one part which is equivalent to the component in OPLS (*Orthogonal in X (OPLS)*) is a matrix effect that needs to be modeled to achieve the best possible prediction and interpretation. The second part (*Orthogonal in X (PCA)*) consists of structured variation that does not affect the prediction but can be interesting to study to further improve the interpretation of complex multivariate data. The X-unrelated variation in Y may be modeled in a similar way.

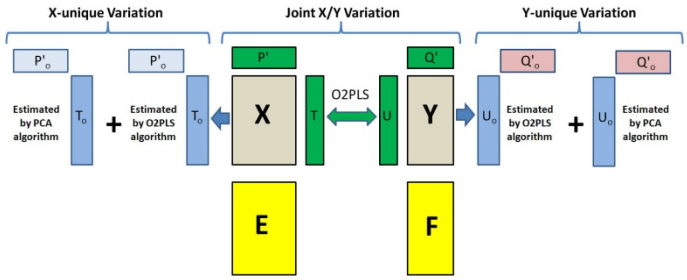


Figure 1**.** Overview of the O2PLS model relating two data tables to each other. Unique variation in X also named variation in X orthogonal to Y in the left-hand side of the Figure. The X/Y joint predictive variation (middle part of the Figure) describes the predictive variation between X and Y, the information overlap. The Unique variation in Y also named variation Orthogonal in Y in the right-hand side of the figure.

The ability to interpret the X/Y joint predictive variation separated from the non-correlated variation implies that the model interpretation is refined and simplified. Furthermore, it should be noted that for the single-y case the OPLS and O2PLS methods are identical. For such a model there can only be one predictive component expressing the joint X/Y predictive variation [4,5].

Combination with the hierarchical approach

Because of its ability to divide the information in X and Y into different parts, the O2PLS model is ideally suited for combining with the hierarchical modeling approach. This is of interest when working

with multi-block process data and spectroscopic data [6,7]. The flexible structure of the O2PLS method also allows setting it up as a filter for peeling off any undesired systematic variability in your data.

Y-Related Profiles

The **Y-Related Profiles** are coefficients rotated displaying the pure profiles of the underlying constituents in X using the assumption of additive Y-variables.

The estimation includes a linear transformation of the coefficient matrix, $Bp(Bp^{T}Bp)^{-1}$, using only the predictive components to compute the coefficients (i.e., the components orthogonal to Y are not included in the computation of the Y related profile).

References OPLS and O2PLS

Trygg, J., and Wold, S., (2002), *Orthogonal Projections to Latent Structures (OPLS)*, Journal of Chemometrics, 16, 119-128.

Trygg, J., (2004), Prediction and Spectral Profile Estimation in Multivariate Calibration, Journal of Chemometrics, 18, 166-172.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, M., and Wold, S., Multi- and Megavariate Data Analysis, Part II, Method Extensions and Advanced Applications, Chapter 23, Umetrics Academy, 2005.

Trygg, J., (2002), O2-PLS for Qualitative and Quantitative Analysis in Multivariate Calibration, Journal of Chemometrics, 16, 283-293.

Trygg, J., and Wold, S., (2003), O2-PLS, a Two-Block (X-Y) Latent Variable Regression (LVR) Method With an Integral OSC Filter, Journal of Chemometrics, 17, 53-64.

Gabrielsson, J., Jonsson, H., Airiau, C., Schmidt, B., Escott, R., and Trygg, J., (2006), *The OPLS methodology for analysis of multi-block batch process data,* Journal of Chemometrics, 20, 362-369.

Eriksson, L., Dyrby, M., Trygg, J., and Wold, S., (2006), *Separating Y-predictive and Y-orthogonal variation in multi-block spectral data*, Journal of Chemometrics, 20, 352-361.

Cluster Analysis (CA), dendrograms, Hierarchical CA (HCA), PLS-Tree

Large datasets are often clustered, grouped, and relationships between variables are often different in different groups (clusters). Cluster analysis tries to find a natural grouping (clustering) of a data set so that there is less variation (greater similarity) within the clusters, and more variation (less similarity) between the clusters. The difference between on the one hand CA, and on the other hand classification and discriminant analysis (CDA) is that in the latter the classes are *pre-defined* by the user, and each observation in the workset belongs to one of these classes. CA has no predefined classes, but after the analysis a set of clusters has been "found", and each observation belongs to one of these. In HCA and PT the "fineness" of the solution, i.e., number of clusters and the observation assignment, can be modified by moving the cursor up and down in the ***dendrogram*** (tree diagram).

CA is particularly useful in the analysis of large data sets, often called "Data Mining". The resulting groups, clusters, can lead to the generation of new ideas – exploratory data analysis – and often better models for parts of the data.

SIMCA has two approaches of hierarchical clustering.

HCA = hierarchical clustering, available on the **Analyze** tab, in the **Clustering** group, by clicking **HCA**, and

PT = PLS-Trees which is a new PLS-based clustering method developed by Umetrics (PLS-TREE™), available on the **Analyze** tab, in the **Clustering** group, by clicking **PLS-Tree**.

Note that much of what is described regarding dendrograms and coloring applies both to HCA and PLS-Trees. Hence, the reader is recommended to also read the following topics and sections:

Hierarchical Cluster Analysis - HCA later in this section.

HCA in the **Analyze** chapter.

PLS-Trees later in this section.

PLS-Tree in the **Analyze** chapter.

Hierarchical Cluster Analysis - HCA

In Hierarchical Clustering Analysis (HCA) a similarity or distance criterion is first specified by the user (default = "Ward", other option: Single Linkage). Then HCA basically starts with as many clusters as there are observations (N). The two closest clusters or observation points are merged, thereafter the two closest clusters or points are again merged, etc., until only one cluster remains. The result is shown by means of a ***dendrogram***. Open plots, e.g., score plots, are colored according to the clusters marked in the dendrogram. For more, see the **HCA** section in the **Analyze** chapter.

Ward clustering

In **Ward** clustering the distance measure is a function that measures the error increase of the model when a pair of clusters are merged into one new cluster. The error function used in SIMCA is the classical error function which calculates the difference in the sum of sum of squares around the mean of each cluster before and after merging two clusters.

This type of clustering works well when the clusters are spherical.

Single linkage

In **single linkage** clustering, the distance between a pair of clusters is the Euclidean distance between the two observations (one in each cluster) closest to each other. This type of clustering works well when the clusters are warped filaments.

PLS-Trees

PLS-Trees is an approach similar to regression and classification trees [3], but using the first score vectors of PLS and PLS-DA instead of the original X-variables for splitting clusters into two.

Thus, the basic idea of a PLS-Tree is to start with a PLS model of the whole data set (the user must before specify what is X and Y, scaling, etc.), and then split the data along the sorted 1st score (**t**) of the PLS model. This split is made to optimize a criterion with three parts:

improving the variance of the score (**t**) so that when combined, the variance in the resulting two sub-groups is as small as possible, and

analogously for the variance of Y, and

do this so that the sizes of the two sub-groups are as equal as possible. Each of these three criteria is in advance given a relative weight by the user, and the criterion to be optimized is the combination of the three parts weighted by the user supplied weights. The user also specifies a minimal size of a cluster beyond which it cannot be further split.

The weight allows the user to focus on the clustering in the X-space, or in the Y-space, or in both, and to more or less focus on splitting into groups of approximately the same size.

Once the split is made and the criterion indicates an improvement, the split goes on with the two sub-groups, and thereafter with the resulting sub-groups of these, until all "branches" have been terminated due to either lack of improvement by further splitting, or that the clusters are reaching the minimal size.

The result is a hierarchical set of PLS models in a tree structure. Each model is an ordinary PLS model which can be further modified, investigated, displayed, and interpreted.

See also the **PLS-Tree** and **HCA - Dendrogram** sections in the **Analyze** chapter for details about coloring etc.

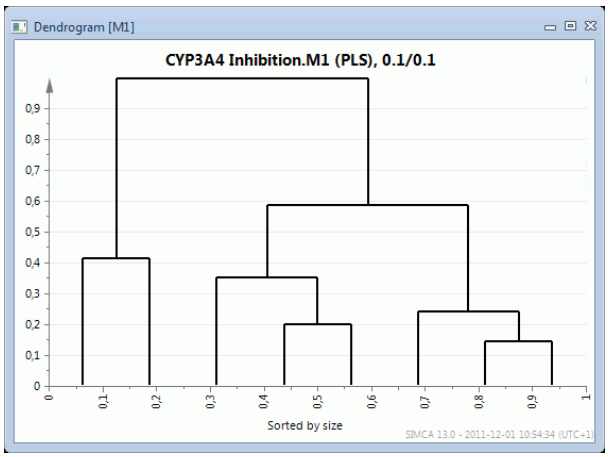


***Figure 1**. A typical PLS-Tree. Each branch of the tree corresponds to a PLS model fitted to a sub-set of observations.*

Calculation of PLS-Trees

Cross validation is used to terminate the branches of the tree, and to determine the number of components of each cluster PLS model.

To accomplish a PLS-Tree the data matrices, X and Y, are first centered and scaled, as usual. A PLS analysis is made of X and Y. Thereafter, the first X-score, t1, is used as the dividing coordinate together with the Y-data (X and Y are sorted along t1). The point on t1 is searched that divides X and Y in two parts, 1 and 2, such that the following expression is minimized:

$$\beta * (N_1 - N_2)^2/(N_1 + N_2)^2 + (1 - \beta) * [\alpha * ((V_{Y1} + V_{Y2})/ V_Y) + (1 - \alpha) * + ((V_{t1} + V_{t2})/V_t)]$$

In the expression above, V denotes variance,

α and β, A and B in SIMCA, are two adjustable parameters.

The parameters A and B both run between 0 and 1. They regulate how the PLS models are split (i.e., how the observations from one upper level PLS model are distributed among two lower level models) according to the score t1, the Y-variable(s) or the group size. The first parameter, **A**, sets the balance between the score t1 and the Y; the **closer to zero** the more weight is attributed to the **score t1**. The second parameter, **B**, takes into account the group size of the resulting clusters; the **closer to zero** the less important it becomes to have **equal group sizes** in the dendrogram. In summary, this means that a division along t1 is sought that minimizes the within groups variation and hence maximizes the between group differences in t1 and Y.

The approach results in the row-wise splitting of data into a tree structure (dendrogram) of PLS models, one split for each cluster (node in the dendrogram). The dendrogram with the associated PLS models is called a PLS-Tree. When Y comprises a discrete matrix with 1/0 columns corresponding to a number of predefined classes, the result is a PLS classification tree.

References cluster analysis

Kriegl, J.M., Eriksson, L., Arnhold, T., Beck, B., Johansson, E., and Fox, T., (2005), *Multivariate Modeling of Cytochrome P450 3A4 Inhibition*, European Journal of Pharmaceutical Sciences, 24, 451-463.

Eriksson, L., Johansson, E., Müller, M., and Wold, S., (2000c), *On the Selection of Training Set in Environmental QSAR When Compounds are Clustered*, Journal of Chemometrics, 14, 599-616.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., (1984) *Classification and Regression Trees,* Wadsworth & Brooks / Cole Advanced Books & Software, Monterey, CA.

Everitt, B.S., Landau, S., Leese, M., (2001), *Cluster Analysis*, Fourth Edition Arnold Publishers, London.

Ward, J.H., (1963), *Hierarchical grouping to optimize an objective function*. J. Am. Stat. Assoc., 58, 236-244.

Vectors available in SIMCA

Modeling using PCA, PLS, OPLS, and O2PLS results in a number of vectors describing the model and its properties, and the residuals.

In the subsections that follow all vectors available in SIMCA are described. The vectors are divided in groups according to their data type in the **Plot/List** dialogs.

For vectors such as variables, scores, and distance to model, see the **Variables and Scores vectors** subsection.

For vectors such as observations, loadings, and Q2VX, see the **Observations and Loadings vectors** subsection.

For vectors pertaining to a component such as Q2, R2X, or SDt see the **Function of component vectors** subsection.

For lag vectors for models with lags defined in the workset, see the **Function of lags vectors** subsection.

For aligned vectors for batch evolution models, see the **Aligned Vectors** subsection.

For Out Of Control vectors for models in batch evolution models, see the **Batch Vectors - Out Of Control Summary** subsection.

Some of the vectors are described with formulas in the **Formulas and descriptions** section later in this chapter.

Note: All vectors available for PLS are also available (with the same description) for OPLS and O2PLS unless otherwise stated. For OPLS and O2PLS these common vectors generally refer to the predictive side.

OPLS and O2PLS specific vectors are described in the **Orthogonal PLS modeling** subsection earlier in this chapter.

Variables and Scores vectors

In the **Plot/List** tab dialogs, the vectors found in the **Items** box with data type **Variables and Scores** are column vectors with the same shape as variables and scores, i.e., with one element per observation.

See the table for the vectors available and their description in alphabetical order. The rightmost column displays the plot/list/spreadsheet where the vector is available, when applicable, in bold text if the vector is displayed by default.

Vector	Description	Displayed
CVgroups	The cross-validation group that each observation is assigned to.	
Date/Time	Variable in the dataset specified as **Date/Time Variable** in the SIMCA import. For more see the **Formatting variable as Date/Time** subsection in the **SIMCA import** chapter.	All line plots displaying observations, for instance: **Home | Scores | Line**
Date/Time-PS	Predictionset variable specified as **Date/Time Variable** in the SIMCA import. For more see the **Formatting variable as Date/Time** subsection in the **SIMCA import** chapter.	All line plots displaying observations, for instance: **Predict | Score PS| Line**

Vector	Description	Displayed
DModX	Distance to the model in X space (row residual SD), after A components (the selected model dimension), for the observations used to fit the model. If you select component 0, it is the standard deviation of the observations with scaling and centering as specified in the workset, i.e., it is the distance to the origin of the scaled coordinate system. (A) = Absolute distance. (N) = Normalized distance. (M) = Mpow weighted residuals.	**Home | DModX**
DModXPS	Distance to the model in the X space (row residual SD), after A components (the selected dimension), for new observations in the predictionset. Displaying component 0, it is the standard deviation of the observations with scaling as specified in the workset times *v* (correction factor for workset observations, see the **Absolute distance to the model of an observation in the workset** subsection in **Appendix A: Statistics**), i.e., it is the distance to the origin of the scaled coordinate system.	Predict | DModX PS
DModXPS+	Combination of DModXPS and tPS, when the latter is declared different from the workset observations.	**Predict | DModX PS**
DModY	Distance to the model in the Y space (row residual SD) after A components (the selected model dimension) for the observations used to fit the model. If you select component 0, it is the standard deviation of the observations with scaling and centering as specified in the workset.	Line or Column under the DModY header in Home | DModX
DModYPS	Distance to the model in the Y space (row residual SD) after A components (the selected model dimension) for observations in the predictionset. If you select component 0, it is the standard deviation of the observations with scaling and centering as specified in the workset.	Line or Column under the DModY header in Predict | DModXPS
Num	Index number: 1, 2, 3 etc.	All lists displaying observations, for instance **Predict | Y PS | List**
ObsID	Numerical observation identifiers, primary, secondary, batch, or phase.	
OLevX	The leverage is a measure of the influence of a point (observation) on the PC model or the PLS model in the X space. The observations leverages are computed as the diagonal elements of the matrix H^0 after A dimensions. $H^0 = T[T'T]^{-1}T'$.	
OLevY	The leverage is a measure of the influence of a point (observation) on the PLS model in the Y space. The observations leverages are computed as the diagonal elements of the matrix H^y after A dimensions. $H^y = U[U'U]^{-1}U'$.	
ORisk	The observation risk is a sensitivity measure and indicates the "risk" (the influence) of including an observation in the workset. ORisk is based on the Y residual (in a PLS/OPLS/O2PLS model) for a selected observation. It is computed from the difference between the residual standard deviation of the selected Y, when the observation is and is not in the model.	

Vector	Description	Displayed
ORisk(pooled)	The pooled ORisk is analogous to ORisk but is valid across all Y variables in a PLS/OPLS/O2PLS model rather than an individual Y variable. It is computed from the difference between the pooled residual standard deviation of the Y variables, when the observation is and is not in the model.	
PModX	Probability of belonging to the model in the X space, for observations used to fit the model. Component 0 corresponds to a point model, i.e., the center of the coordinate system. Observations with probability of belonging of less than 5% are considered to be non-members, i.e., they are different from the normal observations used to build the model.	
PModY	Probability of belonging to the model in the Y space, for observations used to fit the model. Component 0 corresponds to a point model, i.e., the center of the coordinate system. Observations with probability of belonging of less than 5% are considered to be non-members, i.e., they are different from the normal observations used to build the model.	
PModXPS	Probability of belonging to the model in the X space, for new observations in the predictionset. Component 0 corresponds to a point model, i.e., the center of the coordinate system. Observations with probability of belonging of less than 5% are considered to be non-members, i.e., they are different from the normal observations used to build the model.	**Predict | Prediction List**
PModXPS+	The same as PModXPS but based on DModXPS+ instead of DModXPS.	**Predict | Prediction List** **Predict | Classification List**
PModYPS	Probability of belonging to the model in the Y space, for new observations in the predictionset. Component 0 corresponds to a point model, i.e., the center of the coordinate system. Observations with probability of belonging of less than 5% are considered to be non-members, i.e., they are different from the normal observations used to build the model.	
SerrL	Lower limit of the standard error of the predicted response Y for an observation in the workset.	
SerrLPS	Lower limit of the standard error of the predicted response Y for a new observation in the predictionset. SerrLPS is always in original units, i.e., back transformed when Y was transformed, when displayed in the **Prediction List**.	**Predict | Prediction List**
SerrU	Upper limit of the standard error of the predicted response Y for an observation in the workset.	
SerrUPS	Upper limit of the standard error of the predicted response Y for a new observation in the predictionset. SerrUPS is always in original units, i.e., back transformed when Y was transformed, when displayed in the **Prediction List**.	**Predict | Prediction List**

Vector	Description	Displayed
t	Scores t, one vector for each model dimension, are new variables computed as linear combinations of X. They provide a summary of X that best approximates the variation of X only (PC model), and both approximate X and predict Y (PLS/OPLS/O2PLS model).	**Home | Scores**
T2Range	Hotelling's T2 for the selected range of components. It is a distance measure of how far away an observation is from the center of a model hyperplane. For OPLS/O2PLS the range is locked to using first predictive to last Orthogonal in X.	**Home | Hotelling's T2**
T2RangePS	Predicted Hotelling's T2 for the selected range of components. For OPLS/O2PLS the range is locked to using first predictive to last Orthogonal in X.	**Predict | Hotelling's T2PS**
tcv	X score t for the selected model dimension, computed from the selected cross validation round. For PLS/OPLS/O2PLS tcv contains one value per observation and for PCA one value per observation and cross validation round.	**Analyze | CV Scores**
tcvSE	Jack-knife standard error of the X score t computed from all rounds of cross validation.	
to	Orthogonal X score to of the X-part of the OPLS/O2PLS model, for the selected component. It summarizes the unique X variation, i.e., the X variation orthogonal to Y.	**Home | Scores** **Home | Scores | Orth X**
tocv	Orthogonal X score to from the X-part of the OPLS/O2PLS model, for a selected model dimension, computed from the selected cross validation round.	
toPS	Predicted orthogonal X score to of the X-part of the OPLS/O2PLS model, for the observations in the predictionset	**Predict | Scores PS**
toPScv	Matrix of cross validated predicted scores ToPS for the predictionset. Available for OPLS and O2PLS models.	
tPS	Predicted X score t, for the selected model dimension, for the observations in the predictionset.	**Predict | Score PS**
tPScv	Predicted X score t, for the selected model dimension, computed from the selected cross validation round.	
tPScvSE	Jack-knife standard error of the X score t, for the observations in the predictionset, computed from all rounds of cross validation.	
u	Scores u, one vector for each model dimension, are new variables summarizing Y so as to maximize the correlation with the X scores t.	**Analyze | Inner Relation**
ucv	Y score u for the selected model dimension, computed from the selected cross validation round.	
uo	Orthogonal Y score uo of the Y-part of the OPLS/O2PLS model, for the selected component. It summarizes the unique Y variation, i.e., the Y variation orthogonal to X.	**Home | Scores | Orth Y**
uocv	Orthogonal Y score uo of the Y-part of the OPLS/O2PLS model, computed from the selected cross validation round.	
VarDS	Variable from the selected dataset, in original units. Available after selecting a DS in the **Data** box.	

Vector	Description	Displayed
VarID	Variable Identifier.	
VarPS	X variable from the predictionset. Can be displayed in transformed or scaled units.	
XVar	X variable from the workset. Can be displayed in transformed or scaled units.	
XVarPred	A reconstructed variable from the workset. For PLS and PCA models, an X variable from the workset is reconstructed as X=TP'. For OPLS models XVarPred represents the X-values predicted from the given Y-values.	
XVarPredPS	A reconstructed variable from the predictionset. For PLS and PCA models, an X variable from the predictionset is reconstructed as X=TPS * P'. For OPLS models XVarPredPS represents the X-values predicted from the given Y-values.	
XVarPS	X variable from the predictionset. Can be displayed in transformed or scaled units.	**Predict | Prediction List**
XVarRes	X variable residuals for observations in the workset, in original units. Can be displayed in transformed or scaled units.	
XVarResPS	X variable residuals for observations in the predictionset, in original units. Can be displayed in transformed or scaled units.	
XVarResPSSt	X variable residuals for observations in the predictionset, in standardized units (divided by the residual standard deviation). Can be displayed in transformed or scaled units.	
XVarResSt	X variable residuals for observations in the workset, in standardized units (divided by the residual standard deviation). Can be displayed in transformed or scaled units.	
XVarResYRelated	X variable residuals where the systematic variation orthogonal to Y has been removed. Available for OPLS and O2PLS.	
YPred	Predicted values of Y variables for observations in the workset, in original units, i.e., back-transformed when transformations are present. Can be displayed in transformed or scaled units.	**Home | Observed vs. Predicted**
YPredcv	Predicted values of the fitted Ys for observations in the workset, computed from the cross validation procedure.	
YPredPS	Predicted values for Y variables for observations in the predictionset, in original units, i.e. back transformed when transformations are present. Can be displayed in transformed or scaled units.	**Predict | Y PS** **Predict | Prediction List**
YPredErrcv	Prediction error of the fitted Ys for observations in the workset, computed from the cross validation procedure.	
YPredPSConfInt- **YPredPSConfInt+**	Lower/upper limit for the confidence interval of predicted Ys from the predictionset. The limit is calculated from the cross validation and the confidence level specified in model options.	
YPredPScv	Predicted values of the modeled Ys for observations in the predictionset, computed from the cross validation procedure.	

Vector	Description	Displayed
YPredPScvSE	Jack-knife standard error of the prediction of Y for observations in the predictionset, computed from all rounds of cross validation.	
YVar	Y variable from the workset. Can be displayed in transformed or scaled units.	**Home | Observed vs. Predicted**
YVarPS	Y variable from the predictionset. Can be displayed in transformed or scaled units.	**Predict | Y PS**
YVarRes	Y variable residuals for observations in the workset, in original units. Can be displayed in transformed or scaled units.	
YVarResSt	Y variable residuals for observations in the workset, in standardized units (divided by the residual standard deviation). Can be displayed in transformed or scaled units.	
YVarResPS	Y variable residuals for observations in the predictionset, in original units. Can be displayed in transformed or scaled units.	
YVarResPSSt	Y variable residuals for observations in the predictionset, in standardized units (divided by the residual standard deviation). Can be displayed in transformed or scaled units.	

Observations and Loadings vectors

In the **Plot/List** tab dialogs the vectors found in the **Items** box with data type **Observations and Loadings** are row vectors with the same "shape" as an observation, i.e., with one row per variable.

See the table for the vectors available and their description in alphabetical order. The rightmost column displays the plot/list/spreadsheet where the vector is available, when applicable, in bold text if the vector is displayed by default.

Vector	Description	Displayed
Batch VIP	The Batch Variable Importance plot (Batch VIP) is available for batch level models and displays the overall importance of the variable on the final quality of the batch. With phases, the plot displays the importance of a variable by phase. With a PLS model, the Batch VIP displays the plot for one y-variable at a time, with a column per variable and per selected phase. Note: The Batch VIP is only available for scores batch level datasets.	**Batch | Variable Importance**
c	For every dimension in the PLS model there is a c vector. It contains the Y loading weights used to linearly combine the Y's to form the Y score vector u. This means the c vector actually expresses the correlation between the Y's and the X score vector t.	**Home | Loadings**
c(corr)	Y loading weight c scaled as a correlation coefficient between Y and u.	Home | Loadings **Analyze | Biplot**
ccv	Y loading weight c for a selected model dimension, computed from the selected cross validation round.	
ccvSE	Jack-knife standard error of the Y loading weight c computed from the rounds of cross validation.	

Vector	Description	Displayed
co	Orthogonal Y loading weights co combine the Y variables (first dimension) or the Y residuals (subsequent dimensions) to form the scores Uo. These orthogonal Y loading weights are selected so as to minimize the correlation between Uo and T, thereby indirectly between Uo and X. Available for OPLS and O2PLS models.	
cocv	Orthogonal Y loading weights co from the Y-part of the model, for a selected model dimension, computed from the selected cross validation round. Available for OPLS and O2PLS models.	
Coeff	PLS/OPLS/O2PLS regression coefficients corresponding to the unscaled and uncentered X and Y. This vector is cumulative over all components up to the selected one.	Home | Coefficients
CoeffC	PLS/OPLS/O2PLS regression coefficients corresponding to the unscaled but centered X and unscaled Y. This vector is cumulative over all components up to the selected one.	Home | Coefficients
CoeffCS	PLS/OPLS/O2PLS regression coefficients corresponding to centered and scaled X, and scaled (but uncentered) Y. This vector is cumulative over all components up to the selected one.	**Home | Coefficients**
CoeffCScv	PLS/OPLS/O2PLS regression coefficients corresponding to the centered and scaled X and the scaled (but uncentered) Y computed from selected the cross validation round.	
CoeffCScvSE	Jack-knife standard error of the coefficients CoeffCS computed from all rounds of cross validation.	
CoeffMLR	PLS/OPLS/O2PLS regression coefficients corresponding to the scaled and centered X but unscaled and uncentered Y. This vector is cumulative over all components up to the selected one.	Home | Coefficients
CoeffRot	Rotated PLS/OPLS/O2PLS regression coefficients corresponding to the unscaled and uncentered X and Y. This vector is cumulative over all components up to the selected one.	Home | Coefficients
MPowX	The modeling power of variable X is the fraction of its standard deviation explained by the model after the specified component.	
Num	Index number: 1, 2, 3 etc.	
ObsDS	Observation in the dataset, selected in the **Data box,** in original units.	
ObsPS	Observation in the current predictionset, in original units. There is only one current predictionset at a time although many can be specified.	
p	Loadings of the X-part of the model. With a PCA model, the loadings are the coefficients with which the X variables are combined to form the X scores, t. The loading, p, for a selected PCA dimension, represent the importance of the X variables in that dimension. With a PLS model, p expresses the importance of the variables in approximating X in the selected component.	**Home | Loadings**

Vector	Description	Displayed
p(corr)	X loading p scaled as a correlation coefficient between X and t.	Home | Loadings **Analyze | Biplot**
pc	X loading p and Y loading weight c combined to one vector.	**Home | Loadings**
pc(corr)	X loading p and Y loading weight c scaled as correlation coefficients between X and t (p) and Y and u (c), and combined to one vector.	Home | Loadings **Analyze | Biplot**
pccvSE	Jack-knife standard error of the combined X loading p and Y loading weight c computed from all rounds of cross validation.	
pcv	X loading p for a selected model dimension, computed from the selected cross validation round.	
pcvSE	Jack-knife standard error of the X loading p computed from all rounds of cross validation.	
po	Orthogonal loading po of the X-part of the OPLS/O2PLS model. po expresses the unique variability in X not found in Y, i.e, X variation orthogonal to Y, in the selected component.	**Home | Loadings** **Home | Loadings | Orth X**
po(corr)	Orthogonal loading po of the X-part of the OPLS/O2PLS model, scaled as the correlation coefficient between X and to, in the selected component.	
pocv	Orthogonal loading po of the X-part of the OPLS/O2PLS model, for a selected model dimension, computed from the selected cross validation round.	
poso	Orthogonal loading po of the X-part and the projection of to onto Y, so, combined to one vector. Available for OPLS and O2PLS.	**Home | Loadings** **Home | Loadings | Orth X**
pq	X loading weight p and Y loading weight q combined to one vector. Available for OPLS and O2PLS.	**Home | Loadings** **Home | Loadings | Pred X-Y**
q	Loadings of the Y-part of the OPLS/O2PLS model. q expresses the importance of the variables in approximating Y variation correlated to X, in the selected component. Y variables with large q (positive or negative) are highly correlated with t (and X).	**Home | Loadings** **Home | Loadings | Pred X-Y**
qcv	Y loading q for a selected model dimension, computed from the selected cross validation round. Available for OPLS and O2PLS models.	
Q2VX, Q2VY	Predicted fraction, according to cross validation, of the variation of the X (PCA) and Y variables (PLS/OPLS/O2PLS), for the selected component.	**Home | Component Contribution**
Q2VXcum, Q2VYcum	Cumulative predicted fraction, according to cross validation, of the variation of the X variables (PCA model) or the Y variables (PLS/OPLS/O2PLS model).	**Home | Summary of Fit**
qo	Orthogonal loading qo of the Y-part of the OPLS/O2PLS model. qo expresses the unique variability in Y not found in X, i.e, Y variation orthogonal to X, in the selected component.	**Home | Loadings | Orth Y**
qocv	Orthogonal loading qo of the Y-part of the OPLS/O2PLS model, for a selected model dimension, computed from the selected cross validation round.	

Vector	Description	Displayed
qor	qo and r combined to one vector. Available for OPLS and O2PLS.	**Home | Loadings | Orth Y**
r	R is the projection of uo onto X. R contains non-zero entries when the score matrix Uo is not completely orthogonal to X. The norm of this matrix is usually very small but is used to enhance the predictions of X. Available for OPLS and O2PLS.	**Home | Loadings | Orth Y**
R2VX	Explained fraction of the variation of the X variables, for the selected component.	**Home | Component Contribution**
R2VXAdj	Explained fraction of the variation of the X variables, adjusted for degrees of freedom, for the selected component.	**Home | Component Contribution**
R2VXAdjcum	Cumulative explained fraction of the variation of the X variables, adjusted for degrees of freedom.	**Home | X/Y Overview**
R2VXcum	Cumulative explained fraction of the variation of the X variables.	**Home | X/Y Overview**
R2VY	Explained fraction of the variation of the Y variables, for the selected component.	**Home | Component Contribution**
R2VYAdj	Explained fraction of the variation of the Y variables, adjusted for degrees of freedom, for the selected component.	**Home | Component Contribution**
R2VYAdjcum	Cumulative explained fraction of the variation of the Y variables, adjusted for degrees of freedom.	**Home | X/Y Overview**
R2VYcum	Cumulative explained fraction of the variation of the Y variables.	**Home | X/Y Overview**
RMSEcv	Root Mean Square Error, computed from the selected cross validation round.	**Analyze | RMSECV**
RMSEE	Root Mean Square Error of the Estimation (the fit) for observations in the workset.	
RMSEP	Root Mean Square Error of the Prediction for observations in the predictionset.	**Predict | Y PS | Scatter** **Predict | Y PS | Line**
S2VX	Residual variance of the X variables, after the selected component, scaled as specified in the workset.	
S2VY	Residual variance of the Y variables, after the selected component, scaled as specified in the workset.	
so	So is the projection of to onto Y. So contains non-zero entries when the score matrix To is not completely orthogonal to Y. The norm of this matrix is usually very small but is used to enhance the predictions of Y. Available for OPLS and O2PLS models.	**Home | Loadings | Orth X**
VarID	Numerical variable identifiers, primary or secondary.	All lists displaying variables, for instance **Home | Coefficients | List**
VIP	Variable Influence on the Projection. It provides the influence of every term in the matrix X on all the Y's. Terms with VIP>1 have an above average influence on Y. This vector is cumulative over all components up to the selected one.	**Home | VIP**
VIPcv	VIP computed from the selected cross validation round.	

Vector	Description	Displayed
VIPcvSE	Jack-knife standard error of the VIP computed from all rounds of cross validation.	
w	X loading weight that combine the X variables (first dimension) or the X residuals (subsequent dimensions) to form the scores t. This loading weight is selected so as to maximize the correlation between t and u, thereby indirectly between t and Y. X variables with large w's (positive or negative) are highly correlated with u (and Y).	Home | Loadings
w*	X loading weight that combines the original X variables (not their residuals in contrast to w) to form the scores t. In the first dimension w* is equal to w. w* is related to the correlation between the X variables and the Y scores u. $W^* = W(P'W)^{-1}$ X variables with large w* (positive or negative) are highly correlated with u (and Y).	**Home | Loadings**
w*c	X loading weight w* and Y loading weight c combined to one vector.	**Home | Loadings**
w*ccvSE	Jack-knife standard error of the combined X loading weight w* and Y loading weight c computed from all rounds of cross validation.	
w*cv	X loading weight w*, for a selected model dimension, computed from the selected cross validation round.	
w*cvSE	Jack-knife standard error of the X loading weight w* computed from all rounds of cross validation.	
wcv	X loading weight w, for a selected model dimension, computed from the selected cross validation round.	
wcvSE	Jack-knife standard error of the X loading weight w computed from all rounds of cross validation.	
wo	Orthogonal loading weight wo of the X-part of the OPLS/O2PLS model. It combines the X residuals to form the orthogonal X score to. This loading weight is selected so as to minimize the correlation between to and u, thereby indirectly between to and Y.	
wocv	Orthogonal loading weight wo of the X-part of the OPLS/O2PLS model, for a selected model dimension, computed from the selected cross validation round.	
Xavg	Averages of X variables, in original units. If the variable is transformed, the average is in the transformed metric.	
XObs	X variables for the selected observation in the workset in original units. Can be displayed in transformed or scaled units.	
XObsPred	Reconstructed observations as X=TP' from the workset. Can be displayed in transformed or scaled units.	
XObsPredPS	Reconstructed observations as X=TP' from the predictionset. Can be displayed in transformed or scaled units.	
XObsRes	Residuals of observations (X space) in the workset, in original units. Can be displayed in transformed or scaled units.	

Vector	Description	Displayed
XObsResPS	Residuals of observations (X space) in the predictionset, in original units. Can be displayed in transformed or scaled units.	
Xws	Scaling weights of the X variables.	
Y-Related Profiles	Displays the estimated pure profiles of the underlying constituents in X under the assumption of additive Y-variables. Estimation includes a linear transformation of the Coefficient matrix, $Bp(Bp^TBp)^{-1}$, where Bp is the Coefficient matrix using only the predictive components to compute the Coefficient matrix (i.e., the components orthogonal to Y are not included in the computation of Bp). Available for OPLS and O2PLS models.	**Analyze | Y-Related Profiles**
Yavg	Averages of Y variables, in original units. If the variable is transformed, the average is in the transformed metric.	
YObs	Y variables for the selected observation in the workset in original units. Can be displayed in transformed or scaled units.	
YObsRes	Residuals of observations (Y space) in the workset, in original units. Can be displayed in transformed or scaled units.	
YObsResPS	Residuals of observations (Y space) in the predictionset, in original units. Can be displayed in transformed or scaled units.	
Yws	Scaling weights of the Y variables.	

Function of component vectors

For the **Plot/List** tab dialogs the vectors found in the **Items** box with data type **F(Component)** are available per component. Some statistics apply to the whole matrix; others are for selected variable, for each component.

See the table for the vectors available and their description in alphabetical order. The rightmost column displays the plot/list/spreadsheet where the vector is available, when applicable, in bold text if the vector is displayed by default.

Vector	Description	Displayed
Eig	Eigenvalues of the X matrix.	**View | Model Window**
Iter	Number of iterations of the algorithm till convergence.	**View | Model Window**
Num	Index number: 1, 2, 3 etc.	
Q2	Fraction of the total variation of the X block (PCA) or the Y block (PLS) that can be predicted by each component.	**View | Model Window**
Q2(cum)progression	Cumulative Q2 for the extracted components, showing the progression of cumulative values for each added orthogonal component in the OPLS/O2PLS model, e.g. 1+0, 1+1, 1+2. Available for OPLS/O2PLS models with 1+x+0 components.	**Home | Summary of Fit**
Q2cum	Cumulative Q2 for the extracted components.	**View | Model Window** **Home | Summary of Fit**

Vector	Description	Displayed
Q2VX, Q2VY	Predicted fraction, according to cross validation, of the variation of the X (PCA) and Y variables (PLS/OPLS/O2PLS), for the selected component.	**Home | Summary of Fit | Component Contribution**
Q2VXcum, Q2VYcum	Cumulative predicted fraction, according to cross validation, of the variation of the X variables (PCA model) or the Y variables (PLS/OPLS/O2PLS model).	**Home | Summary of Fit | X/Y Overview**
R2(cum)progression	Cumulative fraction of Y variation, showing the progression of cumulative values for each added orthogonal component in the OPLS/O2PLS model, e.g. 1+0, 1+1, 1+2. Available for OPLS/O2PLS models with 1+x+0 components.	**Home | Summary of Fit**
R2VX	Explained fraction of the variation of the X variables, for the selected component.	**Home | Summary of Fit | Component Contribution**
R2VXAdjcum	Cumulative explained fraction of the variation of the X variables, adjusted for degrees of freedom.	Home | Summary of Fit | X/Y Overview
R2VXcum	Cumulative explained fraction of the variation of the X variables.	**Home | Summary of Fit | X/Y Overview**
R2VY	Explained fraction of the variation of the Y variables, for the selected component.	**Home | Summary of Fit | Component Contribution**
R2VYAdjcum	Cumulative explained fraction of the variation of the Y variables, adjusted for degrees of freedom.	**Home | Summary of Fit | X/Y Overview**
R2VYcum	Cumulative explained fraction of the variation of the Y variables.	**Home | Summary of Fit | X/Y Overview**
R2X	Fraction of the total variation of the X block that can be explained by each component.	**View | Model Window.** **Home | Summary of Fit**
R2XAdj	Explained fraction of the variation of the X block, adjusted for degrees of freedom, for the selected component.	**Home | Summary of Fit**
R2XAdjcum	Cumulative explained fraction of the variation of the X block, adjusted for degrees of freedom.	
R2Xcum	Cumulative explained fraction of the variation of the X block.	**View | Model Window.** **Home | Summary of Fit**
R2Y	Fraction of the total variation of the Y block that can be explained by each component.	**View | Model Window.** **Home | Summary of Fit**
R2YAdj	Explained fraction of the variation of the Y block, adjusted for degrees of freedom, for the selected component.	**Home | Summary of Fit**
R2YAdjcum	Cumulative explained fraction of the variation of the Y block, adjusted for degrees of freedom.	
R2Ycum	Cumulative explained fraction of the variation of the Y block.	**View | Model Window.** **Home | Summary of Fit**
RMSEcv	Root Mean Square Error, computed from the selected cross validation round.	**Analyze | RMSECV**

Vector	Description	Displayed
RMSEcv-progression	Root Mean Square Error showing the progression of RMSEcv values for each added orthogonal component in the OPLS/O2PLS model, e.g. 1+0, 1+1, 1+2. Available for OPLS/O2PLS models with 1+x+0 components.	**Analyze | RMSECV**
S2X	Variance of the X block. For component number A, it is the residual variance of X after component A.	
S2Y	Variance of the Y block. For component number A, it is the residual variance of Y after component A.	
SDt	Standard deviation of the X scores, T.	
SDu	Standard deviation of the Y scores, U.	
SSX	Sum of squares of the X block. For component number A, it is the X residual Sum of Squares after component A.	
SSY	Sum of squares of the Y block. For component number A, it is the Y residual Sum of Squares after component A.	
YPredErrcvSE	Jack-knife standard error of the prediction error of the fitted Ys for observations in the workset, computed from the cross validation rounds.	

Function of lags vectors

For the **Plot/List** tab dialogs vectors found in the **Items** box with data type **F(Lags)** are available as functions of lags.

See the table for the vectors available and their description in alphabetical order.

Vector	Description
CoeffCScvSELag	Jack-knife standard error on the coefficients as a function of lag, computed from all cross validation rounds.
CoeffCSLag	Coefficients (for scaled and centered data) of a lagged variable x, for a selected Y as a function of lags.
Num	Index number: 1, 2, 3 etc.
pLag	X loading p of a lagged variable X, as a function of lags.
wLag	X loading weight w of a lagged variable as a function of lags
w*Lag	X loading weight w* of a lagged variable as a function of lags
VIPLag	VIP of a lagged variable X as a function of lags.

Aligned Vectors

In the **Plot/List** tab dialogs the vectors found in the **Items** box with data type **Aligned Vectors** are available for vectors aligned to median length. The average or standard deviation of the vector can be displayed, and selected batches.

Aligned vectors are available for batch evolution models.

Note: All aligned vectors are named '(Aligned)' in addition to the vector name.

See the table for the vectors available and their description in alphabetical order. The rightmost column displays the plot/list/spreadsheet where the vector is available.

Vector	Description	Displayed
DModX	DModX vector aligned to median length. See DModX description previously.	Batch | DModX BCC.
DModXPS	DModXPS vector aligned to median length. See DModXPS description previously.	Batch | DModX PS BCC.
Num	Index vector: 1, 2, 3 etc.	
t	t vector aligned to median length. See t description previously.	Batch | Scores BCC.
T2Range	T2Range vector aligned to median length. See T2Range description previously.	Batch | Hotelling's T2 BCC.
T2RangePS	T2RangePS vector aligned to median length. See T2RangePS description previously.	Batch | Hotelling's T2PS BCC.
Time/Maturity	Time or Maturity variable determining the end point of a Batch/Phase and used as Y in the observation level models. This variable is used to align Batch/Phase to the median length.	All aligned batch control charts.
tPS	tPS vector aligned to median length. See tPS description previously.	Batch | Scores PS BCC.
XVar	XVar vector aligned to median length. See XVar description previously.	Batch | Variable BCC.
XVarPS	XVarPS vector aligned to median length. See XVarPS description previously.	Batch | Variable PS BCC.
YPred	YPred vector aligned to median length. See YPred description previously.	Batch | Observed vs. Time/Maturity BCC.
YPredPS	YPredPS vector aligned to median length. See YPredPS description previously.	Batch | Observed vs. Time/Maturity PS BCC.
YVar	YVar vector aligned to median length. See YVar description previously.	Batch | Observed vs. Time/Maturity BCC.
YVarPS	YVarPS vector aligned to median length. See YVarPS description previously.	Batch | Observed vs. Time/Maturity PS BCC.

Batch Vectors

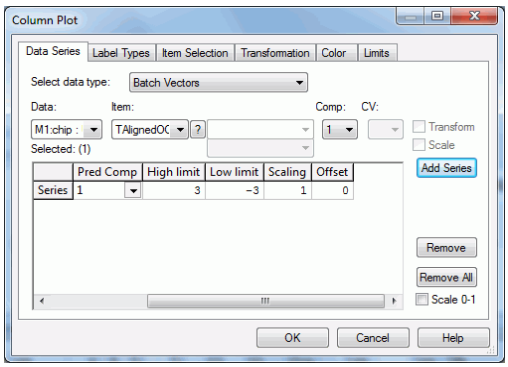
In the **Plot/List** tab dialogs the vectors found in the **Items** box with data type **Batch Vectors** are the Out of Control Summary vectors, **OOCSum**, for all vectors displayed in the batch control charts. That is: DModX, DModXPS, T2Range, T2RangePS, t, tPS, XVar, XVarPS, YPred, and YPredPS. The OOCSum vectors are all named as the mother vector with the suffix **AlignedOOCSum**.

The **OOCSum** for a given vector is the ratio of the area of the vector outside the control limits to the total area inside the limits, for each batch. The **OOCSum** is then expressed as a percent of the area inside the limits. With no part of the batch outside the limits, OOCSum=0.

OOCSum is always computed on aligned vectors, both for workset and predictionset vectors.

The **OOCSum** vector is default displayed using the default limits.

Change the limits as desired under **High limit** and **Low limit** after adding the series.



Vector	Description
DModXAlignedOOCSum	For a plot of DModX vs. Maturity for a batch in the workset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to a DModX vector aligned to median length.
DModXPSAlignedOOCSum	For a plot of DModX vs. Maturity for a batch in the predictionset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area.
T2RangeAlignedOOCSum	For a plot of T2Range vs. Maturity for a batch in the workset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to a T2Range vector aligned to median length.
T2RangePSAlignedOOCSum	For a plot of T2Range vs. Maturity for a batch in the predictionset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to a T2RangePS vector aligned to median length.
TAlignedOOCSum	For a plot of scores (t) vs. Maturity for a batch in the workset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to a t vector aligned to median length.
TPSAlignedOOCSum	For a plot of scores (t) vs. Maturity for a batch in the predictionset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to a tPS vector aligned to median length.
XVarAlignedOOCSum	For a plot of an X-variable vs. Maturity for a batch in the workset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to an XVar vector aligned to median length.
XVarPSAlignedOOCSum	For a plot of an X-variable vs. Maturity for a batch in the predictionset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to an XVarPS vector aligned to median length.
YPredAlignedOOCSum	For a plot of YPred vs. Maturity for a batch in the workset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to an YPred vector aligned to median length.
YPredPSAlignedOOCSum	For a plot of YPred vs. Maturity for a batch in the predictionset the out of control (OOC) sum is the area outside the control limits expressed as a percentage of the total area. The summation is made relative to an YPredPS vector aligned to median length.

Formulas and descriptions

Fitting a PCA, PLS, OPLS, or an O2PLS model generates all the vectors listed in the **Vectors available in SIMCA** section earlier in this chapter. This section describes and provides formulas for a selection of those vectors. The descriptions for PLS vectors are also true for the vectors created when fitting OPLS/O2PLS models unless otherwise stated. When fitting with OPLS/O2PLS a few additional vectors are calculated and these are listed in the same tables.

All vectors that have a 'Y' in the name refer to PLS/OPLS/O2PLS models.

SIMCA will iteratively compute one component at a time, that is: one vector each of X-scores **t**, Y-scores **u** (for PLS/OPLS/O2PLS), weights **w** and **c** (for PLS/OPLS/O2PLS), and loadings **p**.

The components are calculated in descending order of importance.

Prediction vectors

After fitting a model, this model can be used for predictive purposes for a given set of observations.

The following vectors are described in this section:

R2Y, R2X, R2Yadj, R2Xadj

R2V and R2Vadj

Q2 and Q2V

Q2(cum) and Q2V(cum)

Leverages

RSD of observations and variables

Modeling power - MPowX

Hotelling's T2

Missing values correction factor

Score and loading vectors

Distance to the model

Variable importance, VIP

Standard Error on the predicted values

Coefficients

R2Y, R2X, R2Yadj, R2Xadj

R^2Y and R^2X display the fraction of the sum of squares for the selected component, SS explained.

R^2Yadj, R^2Xadj display the variance explained in the model, that is, SS explained corrected for degrees of freedom

R2V and R2Vadj

For every variable in the model, the fraction of SS (R^2V) or variance (R^2Vadj) explained can be displayed. This is computed for both the current component and accumulated over all PC or PLS components. For response variables Y, this corresponds to R^2 (the multiple correlation coefficient), the goodness of fit.

Q2 and Q2V

Q^2: The fraction of the total variation of X or Y that can be predicted by a component, as estimated by cross-validation.

Q^2 is computed as:

$Q^2 = (1.0 - PRESS/SS)$

Q^2V: The fraction of the variation of a variable, e.g. x_k or y_m, that can be predicted by a component, as estimated by cross-validation.

Q^2V is computed as:

$Q^2VX = (1.0 - PRESS/SS)_k$

$Q^2VY = (1.0 - PRESS/SS)_m$

where Q^2VX is available for both PCA and PLS, but Q^2VY only for PLS.

When using SIMCA with cross validation, Q^2 and the limit are displayed in the model window and the significance column informs if the component is significant or not and according to which rule.

For details about the cross validation rules, see the **Cross validation** section later in this chapter.

Q2(cum) and Q2V(cum)

The cumulative Q^2 for the extracted components is computed as:

$Q^2(cum) = (1.0 - \Pi(PRESS/SS)_a)$

where

[a = 1, ...A]

$\Pi(PRESS/SS)_a$ = the product of PRESS/SS for each individual component a.

When a component is insignificant the value *PRESS/SS* is truncated for that component to –0.1, when that value is smaller than –0.1.

The cumulative Q^2V of a variable is similarly computed as:

$Q^2V(cum) = (1.0 - \Pi(PRESS/SS)_{ka})$

where

[a = 1, ...A]

When responses have large values of $Q^2V(cum)$ accumulated over all dimensions, the model for this variable is good. An accumulated value larger than about 0.5 can be considered large.

Leverages

The leverage is a measure of the **influence** of a point (observation) on the PC or PLS model. Leverage (OLev) is proportional to Hotelling's T^2. Observations with high leverages can exert a large influence on the model. A high leverage observation falling near a component axis reinforces the model. A high leverage observation lying far away from a component line causes a rotation of the model.

SIMCA computes the observations leverages in the X and Y spaces as the diagonal elements of the matrices H_0 and H_y, respectively:

$H_0 = T(T'T)^{-1}T'$

$H_Y = U(U'U)^{-1}U'$

RSD of observations and variables

The residual standard deviation (RSD) can be computed for observations and variables. The RSD of an observation in the X or Y space (rows in E and F) is proportional to the observation distance to the hyper plane of the PLS model in the corresponding space (DModX and DModY).

The RSD of an X variable (columns in E) relates to its relevance in the PC or PLS model. The RSD of a Y variable (columns in F) is a measure of how well this response is explained by the PLS model.

Relevance of variables

The importance of a variable in a PC model is indicated by the size of its explained Sum of Squares (R^2X) or variance (R^2Xadj).

A variable is relevant if its modeling power = 1.0. Variables with low modeling power, i.e., around (A/K) are of little relevance (A = number of model dimensions, K= number of variables).

Modeling power - MPowX

The modeling power of a variable is defined as its explained standard deviation:

$Mpow_k = 1.0 - SV_k/SV_{0k}$

where

SV_k = residual standard deviation (RSD) of variable x_k.

SV_{0k}= initial standard deviation of variable k. SV_{0k}= 1.0 if the variable has been autoscaled.

Hotelling's T2

T2Range is basically calculated as the sum over the selected range of components of the scores in square divided by their standard deviations in square, provided the variables were centered. Hence, T2Range is the distance in the model plane (score space) from the origin, in the specified range of components. This means that the same component is selected as 'From' and 'To' components, the plot displays the Hotelling's T2 for that component.

For OPLS and O2PLS the range is locked to 'From' the first predictive 'To' the last orthogonal in x.

The Hotelling's T^2 for observation i, based on A components is:

$T_i^2 = \Sigma ((t_{ia} - t_{avg})^2 / s_{ta}^2)$

where

the summation is done over the range of the selected components,

s^2_{ta} = Variance of t_a according to the class model.

$T_i^2 * (N - A) / A(N - 1)$

is F distributed with A and N-A degrees of freedom.

N = Number of observations in the workset.

A = Number of components in the model or the selected number of components.

Hence if

$T_i^2 > A(N - 1) / (N - A) * F_{critical}$ (p=0.05)

then observation **i** is outside the 95% confidence region of the model.

Ellipse in score plot

The confidence region for a two dimensional score plot of dimension a and b is an ellipse with axis:

$\sqrt{(s^2_{ta\ or\ b} * F(2,N-2,\alpha) * 2 * ((N -1)/(N-2)))}$

The significance level is by default 0.05. On the **File** tab, click **Project Options** and in the **Fit** tab change the default level if desired.

To not display the ellipse, see the **Limits and Regions** and **Limits** subsections in the **Plot and list contextual tabs** chapter.

Missing values correction factor

The NIPALS algorithm has been modified for data containing missing values. A limit has been introduced on the correction factor used with missing values. This stabilizes both the algorithm and the solution. A value of 3.0 for the limit on the correction factor for both the scores and the loadings has generally been found to be appropriate.

Missing values and PCA

A very important problem in industrial applications of PCA, such as multivariate statistical process control applications, is the estimation of scores when the observation vector has missing data. SIMCAs approach to score calculations for such incomplete observations are based on methods described in Nelson, P.R.C., Taylor, P.A., MacGregor, J.F., *Missing Data Methods in PCA and PLS: Score*

Calculations with Incomplete Observation, Chemometrics and Intelligent Laboratry Systems, 35, 45-65, 1996.

SIMCA uses different approaches to compute t and tPS for incomplete observations and if the proportion of missing data is large there can be numerical differences between t and tPS.

Score and loading vectors

The score and loading plots available for plotting are described below.

Note that the score and loading plots complement each other. The position of an observation in a given direction in a score plot is influenced by variables lying in the same direction in the loading plot.

All of these score plots described here will reveal:

Groups

Trends

Outliers

Similarity

T scores (X): t1 vs. t2, ...

T score plots are windows in the X space displaying the observations as situated on the projection plane or hyper plane.

With a 2-dimensonal score plot SIMCA draws the elliptical confidence interval based on Hotelling's T^2.

U scores (Y): u1 vs. u2, ...

U score plots are windows in the Y space, displaying the observations as situated on the projection plane or hyper plane.

T vs. U scores (X&Y): t1 vs. u1, ...

T vs U score plots display the observations in the projected X(T) and Y(U) space, and show how well the Y space correlates to the X space.

P loadings (X): p1 vs. p2,...

P loading plots show the importance of the X variables in the approximation of the X matrix.

W loadings (X): w1 vs. w2, ...

W loading plots display the correlation between the X variables, in the first dimension, or the residuals of the X variables in subsequent dimensions, and the Y or Y residuals scores U(Y).

The **w**'s are the weights that combine the X variables (first dimension) or the residuals of the X variables (subsequent dimensions) to form the scores t. These weights are selected so as to maximize the covariance between T and U, thereby indirectly T and Y.

X variables with large w's (positive or negative) are highly correlated with U(Y).

Variables with large w's are situated far away from the origin (on the positive or negative side) on the plot.

W* loadings (X): w*1 vs. w*2, ...

For every PLS dimension the $\mathbf{W}^*$'s are the weights that combine the original X variables (not their residuals as with w) to form the scores t.

W^* is computed as follows:

$W^* = W(P'W)^{-1}$

From this formula we can see that w^* is equal to w in the first dimension.

Also, when P is equal to W (for all dimensions), W^* is equal to W because W'W (and P'W) is equal to 1.

Relation of W* to the PLS regression coefficients

For a given response Y, the PLS regression coefficients are (for the model Y = XB):

$B = W(P'W)^{-1}C'$

or

B = W*C' and Y = XW*C'

Hence, W^* is directly related to the PLS regression coefficients.

C loadings (Y): c1 vs. c2, ...

C loading plots display the correlation between the Y variables and T(X). The **c**'s are the weights that combine the Y variables with the scores **u**, so as to maximize their correlation with X. Y variables with **large c's** are highly correlated with T(X).

WC or W*C loadings (X&Y): wc1 vs. wc2, ...

WC or W*C loading plots show **both** the X-weights (**w or w***) and Y-weights (**c**), and thereby the correlation structure between X and Y. One sees how the X and Y variables combine in the projections, and how the X variables relate to Y.

TPS scores (X): tPS1 vs. tPS2, ...

The plots On the **Predict** tab classify (new) observations in the predictionset with respect to a PC or PLS model and predicts responses (values of Y variables) for (new) observations (PLS).

For a given set of observations, SIMCA computes the predicted scores: tPS. With missing data in the new observation (denoted by z below), these scores are calculated by means of PLS and the model below, according to Nelson, Taylor, and MacGregor (see **reference list**).

z = P*tPS + e

Distance to the model

This section describes the different distances to the model available in SIMCA for observations in the workset or in the predictionset.

Absolute distance to the model of an observation in the workset

Normalized Distance to the model of an observation in the workset

Distance to the model of new observations in the predictionset

MPow weighted distance to the model

Critical distance to the model

Membership significance level

Absolute distance to the model of an observation in the workset - DModX, DModY(Absolute)

For observations in the workset, SIMCA computes the observation distance to the model in the X space, DModX, and in the Y space, DModY (for PLS models). These distances are equivalent to X and Y residual standard deviations, respectively.

DModX absolute

DModX of an observation in the workset (i.e., that was part of the model) is computed as:

$s_i = \text{sqrt}(\Sigma e_{ik}^2 / (K-A))* v$

The summation is made over the X variables (k) and $\mathbf{e_{ik}}$ are the X-residuals of observation i.

v is a correction factor (function of the number of observations and the number of components) and is slightly larger than one.

This correction factor takes into account the fact that the distance to the model (DModX) is expected to be slightly smaller for an observation that is part of the workset since it has influenced the model.

DModY absolute

DModY of an observation in the workset (i.e., that was part of the model) is computed as:

$s_i = \text{sqrt}(\Sigma f_{im}^2 / M)$

The summation is made over the Y variables (m) and f_{im} are the Y residuals of observation i.

DModY absolute for OPLS/O2PLS

DModY of an observation in the workset, when the fit method is OPLS/O2PLS is computed as:

$s_i = \text{sqrt}(\Sigma f_{im}^2 / M) * v$

The summation is made over the Y variables (m) and f_{im} are the Y residuals of observation i.

Normalized distance to the model of an observation in the workset - DModX, DModY (Normalized)

The normalized distance to the model is the observation absolute DModX or DModY divided by the pooled RSD of the model, s_0, in the X space and in the Y space.

The pooled RSD for the X space is calculated as:

$s_0 = \sqrt{(\Sigma\Sigma e_{ik}^2 / ((N-A-A_0)*(K-A)))}$

$\mathbf{A_0}$ = 1 if model is centered, 0 otherwise.

DModX normalized

DModX normalized = s_i/s_0

where

$\mathbf{s_i}$ is the absolute DModX.

$\mathbf{s_0}$ is the pooled RSD for the X space.

The pooled RSD for the Y space is calculated as:

$s_0 = \sqrt{(\Sigma\Sigma f_{im}^2 / ((N-A-A_0)*M))}$

DModY normalized

DModY normalized = s_i/s_0

where

$\mathbf{s_i}$ is the absolute DModY.

$\mathbf{s_0}$ is the pooled RSD for the Y space.

Absolute distance to the model of new observations in the predictionset - DModXPS, DModYPS (Absolute)

The absolute distance to the model DModXPS of a new observation is computed as the observation RSD:

$s_i = \sqrt{\text{sqrt}(\Sigma e_{ik}^2 / (K-A))}$

The difference in the formula in comparison with the calculation of DModX is that the correction factor in not present.

The distance to the model in the Y space, DModYPS, of a new observation (i.e., that was not part of the model) is computed in the same way as DModY, that is, for an observation that was part of the model:

$s_i = \sqrt{(\Sigma f_{im}^2 / M)}$

Normalized distance of a new observation to the model - DModXPS, DModYPS (Normalized)

The normalized distance to the model is the observation absolute DModXPS or DModYPS divided by the pooled RSD of the model (not the predictionset) s_0:

Normalized DModXPS = s_i/s_0

where

$s_0 = \sqrt{(\Sigma\Sigma e_{ik}^2 / ((N-A-A_0)*(K-A))}$

Normalized DModYPS = s_i/s_0

where

$s_0 = \sqrt{(\Sigma\Sigma f_{im}^2 / ((N-A-A_0)*M)}$

A_0 = 1 if model is centered, 0 otherwise.

Distance to the model augmented - DModXPS+ (Absolute) or (Normalized)

With predictions, the distance to the model can be augmented with a term measuring how far outside (d) the acceptable model domain the projection of the observation falls. The result is called DModXPS+. Hence, for an observation in the predictionset (hence the "PS" label), this is:

$DModXPS+ = [DModXPS^2 + d^2]^{1/2}$

Where

$d = \sum d0_i^2 * S2X_i / (SDt(i))^2$ for components i = 1 to A.

d0 is the distance between the score value t and the score limit for component i for an observation in the predictionset. If t is within the limits, d0 is zero. The score limits are:

Upper limit: Max(t) + SDt/2

Lower Limit: Min(t) - SDt/2

$S2X_i$ is the residual variance.

Here the distance from the projected point of the observation to the model domain is re-expressed in the same units as the residuals by multiplication by the RSD of the workset (s0) and division by the pooled standard deviation of the score vectors in which the projected point is outside the acceptable interval.

The DModXPS+ can be expressed in absolute or normalized, and weighted by the modeling power, just like DModXPS.

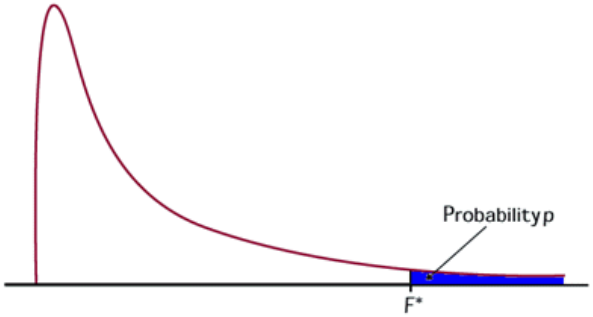
MPow weighted distance to the model - DModX, DModXPS, DModXPS+ (Abs/Norm and weighted)

With MPow weighted DModX/DModXPS/DModXPS+, that is, distance to the model for the X-block weighted by the modeling power, the observation RSD is computed with weighted residuals, where the weights are the modeling powers of the variables.

Critical distance to the model

The critical limit for the distance to model measure (DCrit) is calculated by an inverse cumulative F-distribution function. The three input parameters for this function are the significance level and the degrees of freedom of the numerator and the denominator of the F-distribution. In this case, the numerator is the degrees of freedom of the observations in the training set (DF_obs) and the denominator is the degrees of freedom of the model (DF_mod). DCrit is equal to the value F* in the figure below where the blue area is the probability p that an observation will be outside the DCrit limit. For example, if the significance level is set to 0.95, typically 95% of the observations would have a

DModX value below DCrit. The exact shape and scale of the curve is determined by the degrees of freedom. The F* value can be extracted from statistical tables of the F-distribution.



In SIMCA, the value of the degrees of freedom for the model is calculated as:

DF_mod = √((N-C-A)*(K-A))

where

N = number of observations

C = 1 if all X variables are centered, else 0

A = number of components

K = number of X variables

(*DF_mod* cannot be smaller than 1)

The value of the degrees of freedom for the observations is calculated as:

If K > DF_mod

DF_obs = (M + Q – A) / CORR

else

DF_obs = (M – A) / CORR

where

M = the smallest value of *[K, 100, DF_mod]*

Q = square root of (K-DF_mod)

CORR = N / (N-C-A)

(*N-C-A* and *DF_obs* cannot be smaller than 1).

The expressions above result in a normalized DCrit value. If the critical limit is to be presented in absolute values, DCrit is multiplied by the residual variance in X.

Membership significance level

As default, observations with a probability of class membership less than 5% are considered non-members.

Variable importance, VIP

SIMCA computes the influence on Y of every term (x_k) in the model, called VIP (variable importance in the projection). VIP is the sum over all model dimensions of the contributions VIN (variable influence). For a given PLS dimension, **a**, $(VIN)_{ak}^2$ is equal to the squared PLS weight $(w_{ak})^2$ of that term, multiplied by the explained SS of that PLS dimension.

The accumulated (over all PLS dimensions) value,

$VIP_{ak}^2 = \Sigma(VIN)_k^2$

where the summation is made over a = 1 to A.

This value is then divided by the total explained SS by the PLS model and multiplied by the number of terms in the model.

The final VIP is the square root of that number.

The formula can also be expressed as:

$VIP_{Ak} = \sqrt{(\Sigma(w_{ak}^2 * (SSY_{a-1} - SSY_k))* (K/(SSY_a - SSY_A)))}$

The Sum of squares of all VIP's is equal to the number of terms in the model hence the average VIP is equal to 1.

One can compare the VIP of one term to the others. Terms with large VIP, larger than 1, are the most relevant for explaining Y.

Standard Error and Jack-knifing uncertainties

SIMCA computes standard errors on the predicted Y values in two different ways:

According to Höskuldsson resulting in the Serr vectors.

From Jack-knifing resulting in error bars in plots.

Standard error according to Höskuldsson - Serr

The variance of the prediction (Yhat) for a given response y at a point X_0 according to **Höskuldsson**, is computed as:

$V(Yhat) = (1/N + t_0 (T'T)^{-1} t_0')\sigma^2$

where

t_0 are the scores t of observation X_0

σ^2 is the y error variance estimated from the Sum of Squares of the residuals of y divided by the degrees of freedom for PLS, i.e. $(N-A-A_0)$.

A = number of PLS components and A_0 is 1 or 0, depending on whether y is centered or not.

Confidence intervals can be computed from the standard error by multiplying by a t (distribution) value with the appropriate degrees of freedom.

Jack-knifing (JK) uncertainties

Jack-knifing is a method for finding the precision of an estimate, by iteratively keeping out parts of the underlying data, making estimates from the subsets and comparing these estimates.

In both PCA and PLS, the set of multiple models resulting from the cross-validation is used to calculate jack-knifing uncertainty measures (standard errors and confidence intervals) of scores, loadings, PLS-regression coefficients, predicted Y-values, and VIP. These are displayed by default in column plots, and can also be listed by right-clicking the plot and clicking **Create List**.

Right-clicking a column plot and clicking **Properties**, clicking the **Limits** tab, and then selecting **None** as confidence limit provides the means to remove the confidence intervals if desired.

The standard formula of JK is used (see e.g., B. Efron and G. Gong).

Reference Jack-knifing

Efron, B., and Gong, G., (1983), A Leisurely Look at the Bootstrap, the Jack-knife, and Cross-validation, American Statistician, 37, 36-48.

Coefficients

The PLS model parameters are used to re-express the Y variables as multiple regression models of the X variables:

$Y = XB$

where

X refers to the X matrix, including squared and/or cross terms if those were added

Y is the matrix of responses.

The values of the regression coefficients depend on the scaling of X and Y.

Note: These coefficients are usually **not** independent unless X has been generated by a statistical design.

The types of coefficients available are described in the subsections that follow.

For the OPLS/O2PLS specific **Y Related Profiles**, see the **Y-Related Profiles** subsection earlier in this chapter.

Scaled and centered coefficients - CoeffCS

CoeffCS are the coefficients when X is scaled and centered and Y is scaled. These coefficients are used for interpreting the influence of the variables X on Y.

The scaling weights are those selected in the workset, usually to give the X and Y variables unit variance (autoscaling).

The centered and scaled coefficients are expressed according to the following formula:

$y * ws_m = ybar * ws_m + b_k(x_k - xbar_k) * ws_k + ...$

$... + b_{kk}(z_k^2 * ws_k^2 - m_{kk}) * v_k + ...$

$... + b_{jk}(z_j * z_k * ws_j * ws_k - m_{jk}) + v_{jk} + ...$

where

$z_k = (x_k - xbar_k)$

ws_m = Scaling weights of response y_m

ws_k = Scaling weights of variable x_k

m_{kk} = Average of z_k^2

m_{jk}= Average of $z_j * z_k$

v_k = Scaling weight of z_k^2

v_{ij}= Scaling weight of $z_i * z_j$

Coefficients MLR - CoeffMLR

CoeffMLR are the coefficients when Y is unscaled and uncentered and X is scaled and centered and the second centering and scaling of the squared and cross terms has been removed.

$y = b_0 + b_k(x_k - xbar_k) * ws_k + ...$

$... + b_{kk}(z_k^2 * ws_k^2) + ...$

$... + b_{ik}(z_i * z_k * ws_i * ws_k) + ...$

where

$z_k = (x_k - xbar_k)$

ws_k = Scaling weights of variable x_k

These coefficients correspond to the multiple regression coefficients.

Coefficients unscaled - Coeff

The coefficients when X and Y are unscaled and uncentered are expressed as follows:

$y_m = b_0 + b_1x_1 + \ldots + b_kx_k + b_{11}x_1^2 + b_{12}x_1x_2 + \ldots$

The unscaled coefficients should be used only for computations, for example contour plots, but are difficult to interpret because of the differences in units between the raw variables.

Note: When x was transformed in the model it also needs to be transformed when used in the equation.

Rotated coefficients

When working with spectral data, it is often desirable to estimate the pure spectral profile as they relate to Y.

The PLS coefficients as relating to centered and scaled data are expressed as

$B = W(P'W)^{-1}C'$

The rotated coefficients are obtained by the transformation of the **B** matrix into the pure constituent profile estimates, $\mathbf{K}_{PLS}$.

$K_{PLS} = \mathbf{B}(B^TB)^{-1}$

This corresponds to a projection of the X space into the B space, in order to estimate the pure constituents profile from Y, in other word a direct calibration.

Rotating the coefficients is warranted not only with spectral data, but in any situation, similar to indirect calibration, where Y is assumed to contribute additively to the variation in X.

Reference rotated coefficients

Trygg, J., (2004), Prediction and Spectral Profile Estimation in Multivariate Calibration, Journal of Chemometrics, 18, 166-172.

Coefficients with Y unscaled uncentered, X unscaled centered - CoeffC

The coefficients when Y is unscaled and uncentered, and X is unscaled but centered, are computed as follow:

$y = ybar + b_k(x_k - xbar_k) + ...$

$... + b_{kk}(z_k^2 - m_{kk}) + ...$

$... + b_{jk}(z_j * z_k - m_{jk}) + ...$

where the terms are defined as for the other coefficients.

Note: CoeffC cannot be selected in the **Properties** dialog of the **Coefficient Plot** only in the **Plot/List** tab dialogs.

RMSEE, RMSEP and RMSECV

A popular plot to interpret the performance of a regression model is the scatter plot showing the relationship between the observed Y and the predicted Y. Such a plot can be constructed both for the workset observations and the predictionset observations, provided that the observed Y is known for the latter group.

By using the regression line tool in SIMCA, the resulting R2 of the regression line can be used to quantify the strength of the association between the observed Y and the predicted Y for the workset observations. The closer to unity the stronger the association. The drawback of the obtained association measure, however, is that neither it (R2) nor its residual correspondence (1-R2) relate to the measurement unit of the observed Y.

It is possible to obtain a performance measure that relates to the unit of the observed Y, which is denoted Root Mean Square Error of Estimation (RMSEE). RMSEE is computed as $\sqrt{(\sum(Yobs-Ypred)^2/(N-1-A))}$, where Yobs-Ypred refers to the fitted residuals for the observations in the workset. RMSEE measures the fit of the model. An analogous parameter for the predictionset is denoted Root Mean Square Error of Prediction (RMSEP). RMSEP is computed as $\sqrt{(\sum(Yobs-Ypred)^2/N)}$, where Yobs-Ypred refers to the predicted residuals for the observations in the predictionset. RMSEP measures the predictive power of the model.

An alternative predictivity measure for the model is available through summarizing the cross-validation residuals of the observations in the workset. The predictivity measure obtained is called Root Mean Square Error from cross-validation (RMSECV). RMSECV can be regarded as an intermediary to RMSEE and RMSEP, as it applies to the workset (as does RMSEE) but indicates

predictive power (as does RMSEP). The RMSEE, RMSEP and RMSECV values are listed at the bottom of plots wherever these parameters are applicable.

RMSECV is available through a separate plot in the **Analysis** group on the **Analyze** tab, allowing the evolution of RMSECV across the model components to be displayed.

For OPLS and O2PLS models with 1+x+0 components, the RMSEcv plot is displayed showing the progression. For more see the **RMSEcv plot** subsection in the **Analyze** chapter.

Transform page criteria

In the **Transform** page in the **Workset** dialog the **Min/Max** and **Skewness** values are colored red when a transformation is recommended.

The criteria for the coloring are:

Min/Max is colored red when 0<=Min/Max<0.1.

Skewness is colored red when sTest < -2 or 2 < sTest

where

sTest = Skewness/(√(6.0*dN*(dN-1.0)/((dN-2.0)*(dN+1.0)*(dN+3.0)))))

Skewness is the skewness value listed in the variable list.

dN = number of non missing values in the variable and has to be equal to or larger than 3.

Automatic transformation criteria

When using the **Transform** button to automatically log transform when appropriate, the criteria used to decide if a log transform is needed is the same criteria as used for coloring Skewness and Min/Max red, with the following additions:

When calculating **Min/Max** the Min value used is the smallest not counting 0.

The number of non-missing values has to be larger than 4 and larger than 0.1*number of rows in the dataset.

A small constant is added to the values when needed for the transformation.

Scaling

The results of both PC and PLS modeling are scale dependent. Selecting the scaling of the variables is therefore an important step in fitting projection models such as PC or PLS.

The SIMCA default for all selected variables is a unit variance base weight and a block scaling factor of 1 (no block scaling). For more, see the **Scaling variables** section in the **Home** chapter.

This section describes the computation of autoscaling, block scaling, the scaling weight, the scaling of expanded terms, transformed variables, lagged variables, variables in classes, scaling after reselecting observations, and calculation of the scaling weight.

Autoscaling

If one has no prior information about the importance of the variables, autoscaling all variables to unit variance (UV) is recommended. This is equivalent to making all of the variable axes have the same length, which is, giving all of the variables equal importance.

If one has prior information about the importance of the variables, it may be desirable to scale some variables up or down, by modifying their unit variance weights.

Block scaling

Block-wise scaling is warranted when a data table contains several types (blocks) of variables, with a highly different number of variables in each block. Block-wise scaling allows each block to be thought of as a unit and to be given the appropriate variance which is smaller than if each variable was autoscaled to unit variance.

In such cases you can select to scale a block, i.e., a number of variables of the same type, such that the whole block has

√(Kblock)

or

√(√(Kblock))

variance, where Kblock is the number of variables in the block.

Scaling weight calculation

The final scaling weight (Xws or Yws) is the product of:

Base weight, usually Unit Variance.

Block scaling weight.

Modifier (default = 1; used to scale variables up or down relative to the base scaling weight).

When the variable is not blocked (block scaling weight = 1) and when the modifier is equal to 1, the scaling weight is equal to the base weight.

Scaling of expanded terms

Expanded variables, i.e. squares, cross or cubic terms are displayed in the **Scale** page. You can block scale expanded terms and/or change their modifier, but you cannot change the base type as it is inherited from the mother variable.

Note: Whenever you change the base type of the mother variables, the scaling of both lagged variables and expanded terms is re-computed accordingly.

Computation

To compute expanded variables, all mother variables are first always centered (even if mother variables in themselves are uncentered) and then scaled as specified by the base type of the mother variables.

Scaling and centering of the expanded term after computation

Expanded variable are scaled using the scaling base type of the mother variables.

For cross terms, if mother variables have different scaling types, the expanded term inherits the scaling type of the majority, or for 2 factor interactions the highest ranking scaling type.

The hierarchy of scaling types is as follow:

Centered and then Noncentered.

UV, Pareto, and then None.

Scaling weights for transformed variables

When variables are transformed, the scaling weights are those of the transformed variables.

If you later further transform a variable, the scaling weight is re-computed according to the base type.

If you transform a variable after selecting base type **Freeze**, the scaling weight is not recomputed. It is frozen with the status of the variable and the current selection of observations at the time you selected **Freeze** as base type.

Scaling of variables with classes

When you have grouped observations in classes, and fitted a PC or PLS class model, the scaling weights of the variables are computed using the observations in the class. If a variable has zero variance within the class observations, this variable is not excluded but is given a scaling weight based on the observations of the whole workset.

Scaling of lagged variables

The lagged variables, by default, inherit the scaling weight (and transformation) of the mother variables. It is **strongly recommended not to change the scaling weight of lagged variables.**

Scaling after changing the observation selection

The scaling weights of all variables, except those with base type **Frozen**, are recomputed when you change the selection of observation.

The scaling weights of variables with a frozen base type are not computed when the selection of observations is changed after **Freeze** was selected as base type.

Cross validation

Cross validation for PCA

The cross validation works as follows:

1. Parts of the X data are kept out of model development.
   SIMCA uses the approach of Krzanowski [1] where in two sub rounds, data are first kept out observation-wise (row-wise) to get a set of loading vectors, and second data are kept out variable-wise (column-wise) to get a set of score vectors.
2. The kept out parts are then predicted by the model.
3. The predictions of the kept out parts are compared with the actual values.
4. 1-3 is repeated until all parts have been kept out once and only once.

PRediction Error Sum of Squares - PRESS

The prediction error sum of squares (PRESS) is the squared differences between observed and predicted values for the data kept out of the model fitting. The prediction of the (i, k) element in scaled and centered form is the i^{th} score value multiplied by the k^{th} loading value, where both have been estimated in a CV round when this element was kept out.

This procedure is repeated several times until every data element has been kept out once and only once. The final PRESS then has contributions from all data.

For each component consecutively, SIMCA computes the overall PRESS/SS, where SS is the residual sum of squares of the previous component. SIMCA also computes $(PRESS/SS)_k$ for each X variable (x_k).

Data kept out observation and variable wise

A special (proprietary) PC estimation is used in the CV rounds to minimize the tendency for the present component to partly rotate into later components.

Cross validation for PLS, OPLS and O2PLS

The cross validation works as follows:

1. Rows of the X/Y-data are kept out of model development
2. The kept out parts are then predicted by the model.
3. The predictions of the kept out parts are compared with the actual values.
4. 1-3 is repeated until all parts have been kept out once and only once.

PRediction Error Sum of Squares - PRESS

The prediction error sum of squares (PRESS) is the squared differences between observed and predicted values for the Y-data kept out of the model fitting. The prediction of the (i, m) element in scaled and centered form is the i^{th} score value multiplied by the m^{th} loading value, where both have been estimated in a CV round when this element was kept out.

This procedure is repeated several times until every data element has been kept out once and only once. The final PRESS then has contribution from all data.

PLS Specific

For each component consecutively, SIMCA computes the overall PRESS/SS, where SS is the residual sum of squares of the previous component. This type of cross validation is usually named partial cross validation. SIMCA also computes $(PRESS/SS)_m$ for each Y variable (x_m).

OPLS and O2PLS Specific

The OPLS and O2PLS cross-validation is performed as follows:

The number of predictive or joint X/Y components is estimated; this number may be adjusted using rule R2 after calculating all predictive and orthogonal components.

The number of orthogonal components in X and Y are decided.

For these steps, rule R1 is used to determine the significance of the components.

For all included components, SIMCA computes the overall Q2=1-PRESS/SS, where SS is the sum of squares of Y. This type of cross validation is called full cross validation. SIMCA also computes $(Q2)_m$ for each Y variable.

The cross validation for the Orthogonal in X(PCA) and the Orthogonal in Y(PCA) is performed as for the regular PCA above.

Reference cross validation

Eastment, H. and Krzanowski, W., Crossvalidatory choice of the number of components from a principal component analysis, Technometrics 24 (1982) 73-77.

Martens, H and Naes, T., Multivariate Calibration, 1989.

Cross validation rules - Significant component

When calculating the PCA orthogonal components of O2PLS, the cross validation rules for PCA apply.

Rule 1: R1

A component is significant according to **Rule 1 when**

$Q^2 >$ Limit

where

Limit = 0 for PLS models with more than 100 observations.

Limit = 0.05 for PLS models with 100 observations or less.

Limit = 0.01 for OPLS and O2PLS.

Limit depends on the number of components for PCA. The limit increases with subsequent components to account for the loss in degrees of freedom.

The significance limit for each component is displayed in the **Model Window**.

Rule 2: R2 for PCA

A component is significant according to **Rule 2 when**

$Q^2V >$ Limit for at least

20% of the x-variables when K > 25.

sqrt(K)*log10(max(10, K-20)) when $K \leq 25$.

K = number of x-variables.

Q^2V is Q^2 for individual variables.

Provided the eigenvalue > 1.5 or K < 30.

Rule 2: R2 for PLS, OPLS and O2PLS

A component is significant according to **Rule 2 when**

Q^2V> Limit for at least

20% of the y-variables when M $\geq$ 25.

sqrt(M) when M < 25.

M = number of y-variables.

Q^2V is Q^2 for individual variables.

Rule 3: U and R5

With PCA when a component is insignificant it is first labeled U (Undecided). If the next component is significant and has a similar eigenvalues (tolerance of 5%) as the previous one, then both component together are considered significant. The U of the undecided component is changed to R5.

Cross validation rules - Non Significant component

The significance column of the model summary for a "non significant" component is marked **NS**, **N3**, or **N4**. When calculating the PCA orthogonal components of O2PLS, the cross validation rules for PCA apply.

N3

The component is not significant when the remaining data for that component have insufficient degrees of freedom. For PCA, SIMCA stops when the number of extracted components is larger than N/2 or K/2 whichever is smaller.

N4

The component is not significant according to rule 4 when the explained variance for X (PC) / Y (PLS) is less than 1% and no variable has more than 2% (PC) / 3% (PLS) explained variance.

NS

The component is not significant according to rule 1 or rule 2 and N3 and N4 do not apply.

NS is only displayed when the component is computed with cross validation.

Note: N3 and N4 are always displayed when applicable, irrespective of whether cross validation was used or not.

Cross validation rules for batches

The cross validation rules for batch evolution models are different from other types of models since the objective here is to extract as much of X as possible.

The cross validation rules are:

Minimum of 3 components, if R2X = 85 % continue if next component R2X > 7% and then next component if R2X > 5%.

As many as needed to reach R2X of 85% not exceeding N/2.

K/2 or 15 comp (K=Number of variables) whichever smaller and stop when R2X of a component is less than 3%.

Rule 1: RB1

As long as 85% has NOT been reached and a component explains more than 5% of R2X, it is considered significant and labeled RB1.

Rule 2: RB2

If a component explains between 3% and 5% of R2X, it is considered significant and labeled RB2.

NB1 - stopping rule

NB1, stopping rule, is the label when the number of component exceeds one of the following:

N/2

K/2

15

NS - Not Significant

The component explains less than 3% of R2X and is considered not significant. The label is NS.

PLS time series analysis

Processes often have a "memory", i.e., the state of the process at time **t** is influenced by the state at earlier times, i.e., **t-1**, **t-2**, **t-3**, ..., etc. To model such processes with "memory", Box-Jenkins time series models, ARMA, and ARIMA models (autoregressive / integrated / moving average), have been widely used in process modeling and control. A time series model is basically formulated as an ordinary regression model, but including also *lagged* variables, possibly in both X and Y:

$Y_{t,\,m} = b_{0m}X_t + b_{1m}X_{t-1} + b_{2m}X_{t-2} + \ldots + c_{1m}Y_{t-1} + c_{2m}Y_{t-2} + \ldots + e_t$

The main problem with this model formulation has been that the predictor variables (X_t, X_{t-1}, etc.) usually are very collinear and also noisy. Multiple regression with its assumptions of independence and preciseness of the predictors is therefore difficult to use for the model estimation.

With PLS, however, collinear and noisy predictor variables do not cause any particular problems, so the model can be directly estimated on the form above. Technically, this is done in SIMCA by creating an expanded predictor matrix, $\mathbf{X}^*$, that contains the appropriate selection of the original variables, plus **lagged** variables, i.e., time shifted one, two, or any selected time units.

The PLS analysis of the relationship between $\mathbf{X}^*$ and the pertinent **Y** is straightforward, and will give the ordinary PLS parameters, X and Y-scores **t** and **u**, weights **w** and **c**, and regression coefficients, **b**. These can be plotted and interpreted in the same way as with ordinary PLS models. Additionally, plots of weights (**w**) and coefficients (**b**) for the different lags of a single variable are of interest for the identification of the most important lag.

In this way time series analysis is incorporated in the same framework of modeling as ordinary data analysis of "non-dynamic" data, reducing the complexity and difficulties involved.

References PLS time series

Efron, B., and Gong, G., (1983), A Leisurely Look at the Bootstrap, the Jack-knife, and Cross-validation, American Statistician, 37, 36-48.

Eastment, H. and Krzanowski, W., (1982), Crossvalidatory choice of the number of components from a principal component analysis, Technometrics 24 73-77.

Höskuldsson, A., PLS regression methods. J. Chemometrics, 2, (1988) 211-228.

Nelson, P.R.C., Taylor, P.A., MacGregor, J.F., Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observation, Chemometrics and Intelligent Laboratry Systems, 35, 45-65, 1996.

CV-ANOVA

CV-ANOVA, ANalysis Of VAriance testing of Cross-Validated predictive residuals, is a diagnostic tool for assessing the reliability of PLS, OPLS and O2PLS models. It is implemented for single-Y and multiple-Y models for the relation X $\rightarrow$ Y. The diagnostic is based on an ANOVA assessment of the cross-validatory (CV) predictive residuals of a PLS, OPLS or O2PLS model. The advantages of using

the CV-residuals are that no extra calculations are needed and that this procedure secures reasonably independent data and variance estimates.

Formally, ANOVA is a method to compare two models by the size of their residuals when fitted to the same data [1, 2, 3, 4]. In the regression context, the two models compared are:

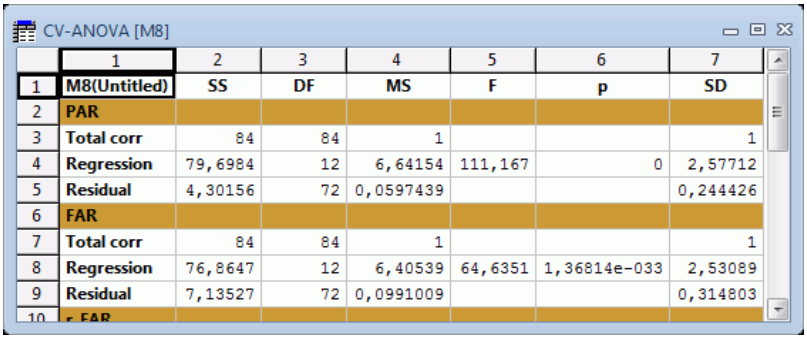
y_i = constant + d_i (1)

y_i = constant + bx_i + e_i (2)

The ANOVA is then made on the size of the sum of squares, SS(d) and SS(e), noting that they are not independent since the data underlying them (y) are the same. In the current context, this means that we test whether the (PLS/OPLS/O2PLS) model has significantly smaller cross validated predictive residuals than just the variation around the global average. In summary, the CV-ANOVA provides a significance test (hypothesis test) of the null hypothesis of equal residuals of the two compared models.

The ANOVA table

The output of the CV-ANOVA is given in a tabulated format using the conventional ANOVA lay-out (see illustration below).


CV-ANOVA [M8]

	1	2	3	4	5	6	7
1	M8(Untitled)	SS	DF	MS	F	p	SD
2	PAR						
3	Total corr	84	84	1			1
4	Regression	79,6984	12	6,64154	111,167	0	2,57712
5	Residual	4,30156	72	0,0597439			0,244426
6	FAR						
7	Total corr	84	84	1			1
8	Regression	76,8647	12	6,40539	64,6351	1,36814e-033	2,53089
9	Residual	7,13527	72	0,0991009			0,314803
10	r FAR						

In the ANOVA table, the following numbers are displayed:

Vector	Type	Description
SS	Total corr (Total corrected)	SS of the Y of the workset corrected for the mean.
	Regression	Fraction of Total Corrected SS accounted for by the model, estimated via the cross validation principle.
	Residual	Difference between Total Corrected and Regression SS, i.e., the fraction of Total Corrected unaccounted for by the model.
DF	Total corr, Regression, residual	The number of degrees of freedom (DF). This is an approximate number based on the experience that PLS needs half the components to reach the same explanation of Y as principal components regression.
MS	Total corr, Regression, residual	By dividing each SS by the respective DF, the corresponding mean squares (MS), or variances, are obtained.
F		The F-test, based on the ratio MS Regression/MS Residual, formally assesses the significance of the model.
p		The p-value indicates the probability level where a model with this F-value may be the result of just chance. The common practise is to interpret a p-value lower than 0.05 as pointing to a significant model.
SD	Standard deviation	Square root of MS.

The great benefit of CV-ANOVA is its user-friendliness. It brings the results of cross-validation to a familiar standard ANOVA format. Thus, CV-ANOVA can be seen as a formal test of the significance of the Q2YCV using the F-distribution. CV-ANOVA is fast and has a minimal additional computation time beyond the one of the standard cross validation. This is in contrast to response permutation testing which is time consuming with large data sets and many components.

References CV-ANOVA

1. Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for experimenters*. Wiley, New York. 2.nd Edition, Wiley 2006.
2. Ståhle, L., and Wold, S., (1989), Analysis of variance (ANOVA), Chemometrics and Intelligent Laboratory Systems, 6, 259-272.
3. Ståhle, L., and Wold, S., (1990), Multivariate analysis of variance (MANOVA), Chemometrics and Intelligent Laboratory Systems, 9, 127-141.
4. Eriksson, L., Trygg, J., and Wold, S., CV-ANOVA for significance testing of PLS and OPLS models, submitted for publication 2008.

Fisher's Exact test

The Fisher Exact test is derived from the probability of the particular classification result and all outcomes more extreme than the one observed. For a 2x2 contingency table the probability of a particular outcome is given by the formula below.

$p = ((a + b)!(c + d)!(a + c)!(b + d)!)/(a! + b! + c! + d! + n!)$

Where ! means factorial (the product of all numbers down to 1).

All probabilities more extreme than the observed pattern are computed and summed to give the probability of the table occurring by chance.

References Fisher's Exact test

1. Fisher, R.A., (1922), On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1):87-94.
2. Fisher, R.A., (1954), Statistical Methods for research workers, Oliver and Boyd.
3. Fisher, R.A., and MacKenzie, W., (1923), Studies in Crop Variation. II. The manurial response of different potato varieties, Journal of Agricultural Science, 13, 311-320.

Control Chart statistics

This section describes the statistics described in the footers of the plots in the subsections:

Nomenclature and notation

Target and Standard deviation

Nomenclature and notation in Control Charts

In the control charts there are a number of statistics displayed in the footer of the plot. This subsection describes the statistics, the data source for the statistics, the standard deviation types calculated, and the model specific statistics.

See the table for the **description of the statistics**.

Statistic	Description
UCL	Upper Control Limit
LCL	Lower Control Limit

Statistic	Description
S	process standard deviation
Target	aim of the process
SAvg	average of standard deviation of subgroup
RAvg	average range of subgroup

In the table the available **data sources** used to estimate S or Target are listed:

There are a number of methods for estimating the standard deviation leading to different **types of standard deviations**. The different standard deviations are described in the table.

Method	Description
SAvg (within)	subgroup standard deviation estimated as the average standard deviation of the subgroups
SAvg (between)	between group standard deviation estimated as the **MSSD**
S	ordinary standard deviation when the sample size is 1
S(Target)	standard deviation estimated around the user entered target.

The **model specific statistics** available are described in the table. Model specific here means that they are calculated using the multivariate model.

Vector	Description
R2X[1]	Fraction of sum of squares for the first component.
Mxx-D-Crit[last component]	critical distance in the DModX plot.
1 – R2X(cum)[last component]	1 – cumulative fraction of sum of squares up to the last component.

Target and Standard deviation

With all control charts, the **Target** and **Standard Dev** (standard deviation) can be **Estimated** from the workset or **User Entered**.

In the plots the following standard deviations and targets are displayed:

S(Mxx) = process standard deviation over all observations in the model.

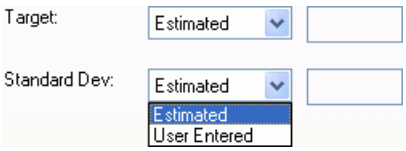
SAvg(within) and **SAvg(between)** see the **Control limits…** sections in the sections for the respective plots. Displayed for >1 subgroups.

S(UE) = user entered standard deviation. Displayed after selecting **User Entered** in the **Standard Dev** box.

Target(Mxx) = overall average when there are no subgroups. With subgroups, the average of the averages of the subgroups.

Target(UE) = user entered target. Displayed instead of Target(Mxx) after selecting **User Entered** in the **Target** box.

***Note**: When displaying plots with items from the predictionset, the estimated target and standard deviation values are computed from the workset of the selected model.*



S-plot

The S-plot is an easy way to visualize an OPLS discriminant analysis model of two classes. It has mainly been used to filter out putative biomarkers from "omics" data e.g. NMR, GC/MS and LC/MS metabolomics data.

In the S-plot both magnitude (intensity) and reliability is visualized. In spectroscopic data the peak magnitude is important as peaks with low magnitude are close to the noise level and thus have a higher risk for spurious correlation. The meaning of high reliability means high effect and lower uncertainty for putative biomarker.

The axes that are plotted in the S-plot from the predictive component are p1 vs p(corr)1, representing the magnitude (modeled covariation) and reliability (modeled correlation) respectively.

Reference S-plot

1. Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., and Trygg, J., (2008), *Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models,* Anal. Chem. 80, 115-122.

***Note**: The vector 'Cov(tp,X)' is named p[1] and the vector 'Corr(tp,X) is named p(corr)[1] in SIMCA.*

Observation Risk

For PLS models, the Observation Risk plot is a measure of the sensitivity of the results of a workset observation in the Y space measured as a change in the Y residuals when the observation is in the workset or not.

Observation Risk is computed for each individual y-variable and for all y-variables together. The column plot displays the observation risk for each y-variable and for the pooled y-variables, as different series.

ORisk = 1 is displayed as a reference line. When ORisk for an observation equals 1, this means that the predicted result for the observation is the same whether the observation is included in the model or not.

ORisk(pooled), the rightmost columns, can be larger than or smaller than the ORisk of individual y-variables. This is due to a correction factor added to prevent DModY, when the observation is in the model, from becoming too small.

A plot displaying observation risk can be created by selecting the vector *ORisk* or *ORisk(pooled)* in one of the standard plots on the **Plot/List** tab.

References

References for MVDA book

Abbas, A., Josefson, M., and Abrahamsson, K., (2011), *Characterization and Mapping of Charotenoids in the Algae Dunaliella and Phaeodactylum using Raman Spectroscopy and Transposed OPLS*, Chemometrics and Intelligent Laboratory Systems, 107, 174-177.

Abbas, A., Josefson, M., Nylund, G.M., Pavia, H., and Abrahamsson K, (2012), *Chemical Images of Marine Bio-active Compounds by Surface Enhanced Raman Spectroscopy and Transposed Orthogonal Partial Least Squares (T-OPLS),* Analytica Chimica Acta, 737, 37-44.

Acar, E., Bro, R., and Schmidt, B., (2008), *New Exploratory Clustering Tool*, Journal of Chemometrics, 22, 19-100.

Aguado, D., Ferrer, A., Ferrer, J., and Seco, A., (2007), *Multivariate SPC of a Sequencing Batch Reactor for Wastewater Treatment*, Chemometrics and Intelligent Laboratory Systems, 85, 82-93.

Albano, C., Dunn, III, W.J., Edlund, U., Johansson, E., Nordén, B., Sjöström, M., and Wold, S., (1978), *Four Levels of Pattern Recognition*, Analytica Chimica Acta, 103, 429-443.

Alsberg, B.K., Woodward, A.M, and Kell, D.B., (1997), *An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach*, Chemometrics and Intelligent Laboratory Systems, 37, 215-239.

Altan, S., Bergum, J., Lindskoug, P., Pfahler, L., Senderak, E., Sethuraman, S., and Vukovinsky, K.E., (2009), *Statistical Considerations in Design Space Development*. Manuscript.

Andersen, S.W., and Runger, G.C., (2011), *Partitioned Partial Least Squares Regression with Application to a Batch Fermentation Process*, Journal of Chemometrics 25, 159-168.

Andersson, G., Kaufmann, P., and Renberg, L, (1996), *Non-Linear Modelling with a Coupled Neural Network – PLS Regression System*, Journal of Chemometrics, 10, 605-614.

Andersson, P., Haglund, P., Tysklind, M., (1997a), *The Internal Barriers of Rotation for the 209 Polychlorinated Biphenyls*, Environmental Science and Pollution Research, 4, 75-81.

Andersson, P., Haglund, P., and Tysklind, M., (1997b), *Ultraviolet Absorption Spectra of all 209 Polychlorinated Biphenyls Evaluated by Principal Component Analysis*, Fresenius Journal of Analytical Chemistry, 357, 1088-1092.

Andersson, G., (1998), *Novel Nonlinear Multivariate Calibration Methods*, Ph.D. Thesis, The Royal Institute of Technology, Stockholm, Sweden.

Andersson, P.M., Sjöström, M., and Lundstedt, T., (1998), *Preprocessing Peptide Sequences for Multivariate Sequence-Property Analysis*, 42, 41-50.

Andersson, P.M., Sjöström, M., Wold, S., and Lundstedt, T., (2001), *Strategies for Subset Selection of Parts of an In-house Chemical Library*, Journal of Chemometrics, 15, 353-369.

Andersson, P.M., and Lundstedt, T., (2002), *Hierarchical Experimental Design Exemplified by QSAR Evaluation of a Chemical Library Directed Towards the Melanocortin 4 Receptor*, Journal of Chemometrics, 16, 490-496.

Andersson, M., Svensson, O., Folestad, S., Josefson, M., and Wahlund, K.G., (2005), *NIR Spectroscopy on Moving Solids Using a Scanning Grating Spectrometer – Impact on Multivariate Process Analysis*, Chemometrics and Intelligent Laboratory Systems, 75, 1-11.

Andersson, M., (2009), *A Comparison of Nine PLS1 Algorithms*, Journal of Chemometrics, 23, 518-529.

Andre, M., (2003), *Multivariate Analysis and Classification of the Chemical Quality of 7-aminocephalosporanic-acid Using Near-infrared Reflectance Spectroscopy*, Analytical Chemistry, 75, 3128-3135.

Anonymous, SIMCA-P manual, Umetrics AB.

Antti, H., Bollard, M.E., Ebbels, T., Keun, H., Lindon, J.C., Nicholson, J.K, and Holmes, E., (2002), *Batch Statistical Processing of 1H-NMR-derived Urinary Spectral Data*, Journal of Chemometrics, 16, 461-468.

Antti, H., Ebbels, T.M.D., Keun. H.C., Bollard, M.A., Beckonert, O., Lindon, J.C., Nicholson, J.K., and Holmes, E., (2004), *Statistical Experimental Design and Partial Least Squares Regression Analysis of Biofluid Metabonomic NMR and Clinical Chemistry Data for Screening of Adverse Drug Effects*, Chemometrics and Intelligent Laboratory Systems, 73, 139-149.

Argyri, A.A., Jarvis, R.M., Wedge, D., Xu, Y., Panagou, E.Z., Goodacre, R., and Nychas, G.J.E., (2913), *A comparison of Raman and FTIR spectroscopy for the prediction of meat spoilage*, Food Control, 29, 461-470.

Atif, U., Earll, M., Eriksson, L., Johansson, E., Lord, P., and Margrett, S., (2002), *Analysis of Gene Expression Datasets Using Partial Least Squares Discriminant Analysis and Principal Component Analysis*. In: Martyn Ford, David Livingstone, John Dearden and Han Van de Waterbeemd (Eds.), Euro QSAR 2002 Designing Drugs and Crop Protectants: processes, problems and solutions. Blackwell Publishing, ISBN 1-4051-2561-0, pp 369-373.

Austel, V., (1995), *Experimental Design*, In: Mannhold, R., Krogsgaard-Larsen, P., and Timmerman, H., (Eds.), Methods and Principles in Medicinal Chemistry, Vol 2, VCH, Weinheim, Germany, pp. 49-62.

Babamoradi, H., van den Berg, F., and Rinnan, Å., (2013), *Bootstrap based confidence limits in prinicpal component analysis – A case study*, Chemometrics and Intelligent Laboratory Systems, 120, 97-105.

Bakshi, B.R., (1999), *Multiscale Analysis and Modelling Using Wavelets*, Journal of Chemometrics, 13, 415-434.

Barnes, R.J., Dhanoa, M.S., and Lister, S.J., (1989), *Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra*, Applied Spectroscopy, 43, 772-777.

Baroni, M., Clementi, S., Cruciani, G., Kettaneh-Wold, S., and Wold, S., (1993a), *D-Optimal Designs in QSAR*, Quantitative Structure-Activity Relationships, 12, 225-231.

Baroni, M., Constatino, G., Cruciani, G., Riganelli, D., Valigi, R., and Clementi, S., (1993b), *Generating Optimal PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems,* Quantitative Structure-Activity Relationships, 12, 9-20.

Barnett, V., and Lewis, T., (1994), *Outliers in Statistical Data*, John Wiley & Sons, Chichester, England.

Bassbasi, M., Hafid, A., Plastikanov, S., Tauler, R., and Oussama, A., (2013), *Study of motor oil adulteration by infrared spectroscopy and chemometrics methods*, Fuel, 104, 798-804.

Barros, A. S., and Rutledge, D.N., (2004), *PLS_Cluster: A Novel Technique for Cluster Analysis*, Chemometrics and Intelligent Laboratory Systems, 70, 99-112.

Berglund, A., and Wold, S., (1997a), *INLR, Implicit Non-Linear Latent Variable Regression*, Journal of Chemometrics, 11, 141-156.

Berglund, A., DeRosa, M.C., and Wold, S., (1997b), *Alignment of Flexible Molecules at Their Receptor Site Using 3D Descriptors and Hi-PCA*, Journal of Computer-Aided Molecular Design, 11, 601-612.

Berglund, A., and Wold, S., (1999), *A Serial Extension of Multi Block PLS*, Journal of Chemometrics, 13, 461-471.

Berglund, A., Kettaneh, N., Uppgård, L.L., Wold, S., Bandwell, N., and Cameron, D.R., (2001), *The GIFI Approach to Non-Linear PLS Modelling*, Journal of Chemometrics, 15, 321-336.

Berntsson, O., Danielsson, L.G., Johansson, M.O., and Folestad, S., (2000), *Quantitative Determination of Content in Binary Powder Mixtures Using Diffuse Reflectance Near Infrared Spectrometry and Multivariate Analysis*, Analytica Chimica Acta, 419, 45-54.

Berntsson, O., (2001), *Characterization and Application of Near Infrared Spectroscopy for Quantitative Process Analysis of Powder Mixtures*, PhD Thesis, KTH, Stockholm, Sweden, ISBN 91-7283-074-3.

Bharati, M.H., Liu, J.J., and MacGregor, J.F., (2004), *Image Texture Analysis: Methods and Comparisons*, Chemometrics and Intelligent Laboratory Systems, 72, 57-71.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., and Pagès, J., (2000), *NIR Calibration in Non-linear Systems: Different PLS Approaches and Artificial Neural Networks*, Chemometrics and Intelligent Laboratory Systems, 50, 75-82.

Blum, D.J.W., and Speece, R.E., (1990), *Determining Chemical Toxicity to Aquatic Species*, Environmental Science and Technology, 24, 284-293.

Bouveresse, D.J.R., Barros, A.S., and Rutledge, D.N., (2007), *Generalised PLS_Cluster: An Extension of PLS_Cluster for Interpretable Hierarchical Clustering of Multivariate Data*, Sensing and Instrumentation for Food Quality and Safety, 1, 79-90.

Box, G.E.P., Hunter, W.G., and Hunter, J.S., (1978), *Statistics for Experimenters*, John Wiley & Sons, Inc., New York.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C., (1994), *Time Series Analysis - Forecasting and Control*, 3rd edition, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., (1984), *Classification and Regression Trees*, Wadsworth & Brooks/Cole, Monterey, CA.

Brereton, R.G., (2009), *Chemometrics for Pattern Recognition*, Wiley & Sons, Inc., New York.

Bro, R., (1996), *Håndbog i Multivariabel Kalibrering*, KVL, Copenhagen, Denmark.

Bro, R., (1999), *Exploratory Study of Sugar Production Using Fluorescence Spectroscopy and Multi-way analysis*, Chemometrics and Intelligent Laboratory System, 46, 133-147.

Bu, D., Wan, B., and McGeorge, G., (2013), *A discussion on the use of prediction uncertainty estimation of NIR data in partial least squares for quantitative pharmaceutical tablet assay methods*, Chemometrics and Intelligent Laboratory Systems, 120, 84-91.

Burnham, A.J., Viveros, R., and MacGregor, J.F., (1996), *Frameworks for Latent Variable Multivariate Regression*, Journal of Chemometrics, 10, 31-45.

Burnham, A.J., MacGregor, J.F, and Viveros, R., (1999), *A Statistical Framework for Multivariate Latent Variable Regression Methods Based on Maximum Likelihood,* Journal of Chemometrics, 13, 49-65.

Burnham, A.J., MacGregor, J., and Viveros, R., (2001), *Interpretation of Regression Coefficients Under a Latent Variable Regression Model*, Journal of Chemometrics, 15, 265-284.

Butz, J., de la Cruz, L., DiTonno, J., DeBoyace, K., Ewing, G., Donovan, B., Medendorp, J., (2011), *Raman Spectroscopy for the Process Analysis of the Manufacturing of a Suspension Metered Dose Inhaler*, Journal of Pharmaceutical and Biomedical Analysis, 54, 1013–1019.

Buydens, L.M.C., Reijmers, T.H., Beckers, M.L.M., and Wehrens, R., (1999), *Molecular Data Mining: a Challenge for Chemometrics*, Chemometrics and Intelligent Laboratory Systems, 49, 121-133.

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J., (2007), *Data Integration in Plant Biology: the O2PLS Method for Combined Modeling of Transcript and Metabolite Data*, The Plant Journal, 52, 1181–1191.

Byrne, T., and Wold, S., (2011), *Data Mining using PLS-Trees and Other Projection Methods*, in: Advanced Semiconductor Manufacturing Conference (ASMC), 22nd Annual IEEE/SEMI, 16-18 May 2011.

Böcker, A., Schneider, G., and Teckentrup, A., (2006), *NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and its Application to Virtual Screening*, Journal Chemical and Information Modeling, 46, 2020-2029.

Börås, L., Sjöström, J., and Gatenholm, P., (1997), *Characterization of Surfaces of CMTP Fibers Using Inverse Gas Chromatography Combined with Multivariate Data Analysis*, Nordic Pulp & Paper Research Journal, 12, 220-224.

Carlson, R., Lundstedt, T., and Albano, C., (1985), *Screening of Suitable Solvents in Organic Synthesis. Strategies for Solvent Selection*, Acta Chemica Scandinavica, B39, 79-91.

Carlson, R., (2004), *Designs For Explorative Experiments in Organic Synthetic Chemistry*, Chemometrics and Intelligent Laboratory Systems, 73, 151-166.

Carlson, R., and Carlson, J.E., (2005), *Design and Optimization in Organic Synthesis – Second Revised and Enlarged Edition*, Data Handling in Science and Technology, Elsevier, ISBN: 0-444-51527-5.

Cederkvist, H.R., Aastveit, A.H., and Naes, T., (2005), *A Comparison of Methods for Testing Differences in Predictive Ability*, Journal of Chemometrics, 19, 500-509.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis R.W., (1998), *A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle*, Molecular Cell, 2, 65-73.

Chorell, E., Pinkner, J.S., Bengtsson, C., Banchelin, T.S.L., Edvinsson, S., Linusson, A., Hultgren, S.J., and Almqvist, F., *Mapping pilicide anti-virulence effect in Escherichia Coli, a comprehensive structure-activity study*, Bioorganic & Medicinal Chemistry, 20, 3128-3142, 2012.

Cloarec, O., Dumas, M.E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E., Nicholson, J., (2005), *Statistical Total Correlation Spectroscopy: an Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets*, Analytical Chemistry, 77, 1282–1289.

Cocchi, M., and Johansson, E., (1993), *Amino Acids Characterization by GRID and Multivariate Data Analysis*, Quantitative Structure-Activity Relationships, 12, 1-8.

Connor, K., Safe, S., Jefcoate, C.R., and Larsen, M., (1995), *Structure-Dependent Induction of CYP2B by Polychlorinated Biphenyl Congeners in Female Sprague-Dawley Rats*, Biochemical Pharmacology, 50, 1913-1917.

Coomans, D., Broeckaert, I., Derde, M.P., Tassin, A., Massart, D.L., and Wold, S., (1984), *Use of a Microcomputer for the Definition of Multivariate Confidence Regions in Medical Diagnosis Based on Clinical Laboratory Profiles*, Computational Biomedical Research, 17, 1-14.

Cronin, M.T.D, and Schultz, T.W., (2003), *Pitfalls in QSAR*, Journal of Molecular Structure (Theochem), 622, 39-51.

Cramer, R.D., Pattersson, D.E., and Bunce, J.D., (1988), *Comparative Molecular Field Analysis (CoMFA). I. Effects of Shape on Binding of Steroids to Carried Proteins*, Journal of American Chemical Society, 110, 5959.

Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R., and Clementi, S., (2004), *Peptide Studies by Means of Principal Properties of Amino Acids Derived from MIF Descriptors*, Journal of Chemometrics, 18, 146-155.

Czekaj, T., Wu, W., and Walczak, B., (2005), *About Kernel Latent Variable Approaches and SVM*, Journal of Chemometrics, 19, 341-354.

Dahl, K.S., Piovoso, M.J., and Kosanovich, K.A., (1999), *Translating Third-order Data Analysis Methods to Chemical Batch Processes*, Chemometrics and Intelligent Laboratory Systems, 46, 161-180.

Davis, O.L., and Goldsmith, P.L., (1986), *Statistical Methods in Research and Production*, Longman, New York.

Dayal, B., MacGregor, J.F., Taylor, P.A., Kildaw, R., and Marcikic, S., (1994), *Application of Feedforward Neural Networks and Partial Least Squares Regression for Modelling Kappa Number in a Continuous Kamyr Digester*, Pulp and Paper Canada, 95, 26-32.

DeAguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L., and Phan-Than-Luu, R., (1995), *D-Optimal Designs*, Chemometrics and Intelligent Laboratory Systems, 30, 199-210.

Dearden, J., (1985), *Partitioning and Lipophilicity in Quantitative Structure-Activity Relationships,* Environmental Health Perspectives, 61, 203-228.

De Jong, S., (1993), *PLS Fits Closer Than PCR*, Journal of Chemometrics, 7, 551-557.

De Jong, S., Wise, B.M., and Ricker, N.L., (2001), *Canonical Partial Least Squares and Continuum Power Regression*, Journal of Chemometrics, 15, 85-100.

Deneer, J.W., Sinninge, T.L., Seinen, W., and Hermens, J.L.M., (1987), *Quantitative Structure-Activity Relationships for the Toxicity and Bioconcentration Factor of Nitrobenzene Derivatives Towards the Guppy (Poecilia reticulata)*, Aquatic Toxicology, 10, 115-129.

Deneer, J.W., van Leeuwen, C.J., Seinen, W., Maas-Diepeveen, J.L., and Hermens, J.L.M., (1989), *QSAR Study of the Toxicity of Nitrobenzene Derivatives Towards Daphnia Magna, Chlorella Pyrenoidosa and Photobacterium Phosphoreum*, Aquatic Toxicology, 15, 83-98.

Denham, M.C., (1997), *Prediction Intervals in Partial Least Squares*, Journal of Chemometrics, 11, 39-52.

Dhanoa, M.S., Lister, S.J., Sanderson, R., and Barnes, R.J., (1994), *The Link Between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformation of NIR Spectra*, Journal of Near Infrared Spectroscopy, 2, 43-47.

Drewry, D.H., and Young, S.S., (1999), *Approaches to the Design of Combinatorial Libraries*, Chemometrics and Intelligent Laboratory Systems, 48, 1-20.

Duarte, I., Barros, A., Belton, P.S., Righelato, R., Spraul, M., Humper, E., and Gil, A.M., (2002), *High-resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the Characterization of Beer*, Journal of Agricultural and Food Chemistry, 50, 2475-2481.

Duchesne, C., and MacGregor, J.F., (2009), *Multivariate Analysis and Optimization of Process Variable Trajectories for Batch Processes*, Chemometrics and Intelligent Laboratory Systems, 51, 125-137.

Dumarey, M., Wikström, H., Fransson, M., Sparén, A. Tajarobi, P., Josefson, M., and Trygg, J., *(2011)*, Combining experimental design and orthogonal projections to latent structures to study the influence of microcrystalline cellulose properties on roll compaction, International Journal of Phramaceutics, 416, 110-119.

Dunn, III, W.J., (1989), *Quantitative Structure-Activity Relationships (QSAR)*, Chemometrics and Intelligent Laboratory Systems*,* 6, 181-190.

Dyrby, M., Petersen, R.V., Larsen, J., Rudolf, B., Nørgaard, L., Engelsen, S.B., (2004), *Towards On-line Monitoring of the Composition of Commercial Carrageenan Powders*, Carbohydrate Polymers, 57, 337-348.

Efron, B., and Gong, G., (1983), *A Leisurely Look at the Bootstrap, the Jack-knife, and Cross-validation*, American Statistician, 37, 36-48.

Elg-Kristoffersson, K., Sjöström, M., Edlund, U., Lindgren, Å., and Dolk, M., (2002) *Reactivity of Dissolving Pulp: Characterisation Using Chemical Properties, NMR Spectroscopy and Multivariate Data Analysis*, Cellulose, 9, 159-170.

Ergon, R., (1998), *Dynamic System Multivariate Calibration by System Identification Methods*, Modelling Identification and Control, 19, 77-97.

Eriksson, L., Jonsson, J., Sjöström, M., Wikström, C., and Wold, S., (1988), *Multivariate Derivation of Descriptive Scales for Monosaccharides*, Acta Chemica Scandinavica, 42, 504-514.

Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Sjöström, M., Wold, S., Sandström, B., and Svensson, I., (1991), *A Strategy for Ranking Environmentally Occurring Chemicals. Part V: The Development of two Genotoxicity QSARs for Halogenated Aliphatics*. Environmental Toxicology and Chemistry, 10, 585-596.

Eriksson, L., Sandström, B.E., Tysklind, M., and Wold, S., (1993), *Modelling the Cytotoxicity of Halogenated Aliphatic Hydrocarbons. Quantitative Structure-Activity Relationships for the IC50 to Human HeLa Cells*, Quantitative Structure-Activity Relationships, 12, 124-131.

Eriksson, L., and Hermens, J.L.M., (1995a), *A Multivariate Approach to Quantitative Structure-Activity and Structure-Property Relationships,* In: The Handbook of Environmental Chemistry, (Ed.) J. Einax, Vol2H, Chemometrics in Environmental Chemistry, Springer Verlag, Berlin.

Eriksson, L., Hermens, J.L.M., Johansson, E., Verhaar, H.J.M. and Wold, S., (1995b), *Multivariate Analysis of Aquatic Toxicity Data with PLS*, Aquatic Sciences, 57, 217-241.

Eriksson, L., and Johansson, E., (1996), *Multivariate Design and Modelling in QSAR*, Chemometrics and Intelligent Laboratory Systems, 34, 1-19.

Eriksson, L., Johansson, E., Müller, M., and Wold, S., (1997), *Cluster-based Design in Environmental QSAR*, Quantitative Structure-Activity Relationships, 16, 383-390.

Eriksson, L., Johansson, E., Tysklind, M., and Wold, S., (1998), *Pre-processing of QSAR Data by Means of Orthogonal Signal Correction*, The QSAR and Modelling Society, Newsletter 1998, www/pharma-ethz.ch./qsar.

Eriksson, L., Andersson, P., Johansson, E., Tysklind, M., Sandberg, M., and Wold, S., (1999a), *The Constrained Principal Property Space in QSAR – Directional and Non-Directional Modelling Approaches*, Proceedings 12th European Symposium on QSAR, August 1998, Copenhagen, Denmark.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., and Wold, S., (1999b), *Design of Experiments – Principles and Applications*, Umetrics AB.

Eriksson, L., Trygg, J., Johansson, E., Bro, R., and Wold, S., (2000a), *Orthogonal Signal Correction, Wavelet Analysis, and Multivariate Calibration of Complicated Process Fluorescence Data*, Analytica Chimica Acta, 420, 181-195.

Eriksson, L., Johansson, E., Lindgren, F., and Wold, S., (2000b), *GIFI-PLS: Modeling of Non-Linearities and Discontinuities in QSAR*, Quantitative Structure-Activity Relationships, 19, 345-355.

Eriksson, L., Johansson, E., Müller, M., and Wold, S., (2000c), *On the Selection of Training Set in Environmental QSAR When Compounds are Clustered*, Journal of Chemometrics, 14, 599-616.

Eriksson, L., Hagberg, P., Johansson, E., Rännar, S., Whelehan, O., Åström, A., and Lindgren, T., (2001), *Multivariate Process Monitoring of a Newsprint Mill. Application to Modelling and Predicting COD Load Resulting from Deinking of Recycled Paper*, Journal of Chemometrics, 15, 337-352.

Eriksson, L., Johansson, E., Lindgren, F., Sjöström, M., and Wold, S., (2002), *Megavariate Analysis of Hierarchical QSAR Data*, Journal of Computer-Aided Molecular Design, 16, 711-726.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., and Gramatica, P., (2003), *Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-based QSAR,* Environmental Health Perspectives, 11, 1361-1375.

Eriksson, L., Arnhold, T., Beck, B., Fox, T., Johansson, E., and Kriegl, J.M., (2004a), *Onion Design and its Application to a Pharmaceutical QSAR Problem*, Journal of Chemometrics, 18, 188-202.

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J., and Wold, S., (2004b), *Using Chemometrics for Navigating in the Large Data Sets of Genomics, Proteomics and Metabonomics*, Analytical and Bioanalytical Chemistry, 380, 419-429.

Eriksson, L., Damborsky, J., Earll, M., Johansson, E., Trygg, J., and Wold, S., (2004c), *Three-block Bi-focal PLS (3BIF-PLS) and its application in QSAR*, SAR and QSAR in Environmental Research, 5/6, 481-499.

Eriksson, L., Gottfries, J., Johansson, E., and Wold, S., (2004d), *Time-resolved QSAR: an Approach to PLS Modelling of Three-way Biological Data*, Chemometrics and Intelligent Laboratory Systems, 73, 73-84.

Eriksson, L., Antti, H., Holmes, E., and Johansson, E., (2005), in: Robertson, D. (Ed.), Modern Safety Assessment Methods of Compounds and Biomarkers: Metabonomics and Evaluation, *Chapter 8: Multi- and Megavariate Data Analysis: Finding and Using Regularities in Metabonomics Data*, Marcel Dekker/CRC Press, ISBN: 0-8247-2665-0.

Eriksson, L., Wold, S., and Trygg, J., (2005), *Multivariate Analysis of Congruent Images (MACI)*, Journal of Chemometrics, 19, 393-403.

Eriksson, L., Trygg, J., and Wold, S., (2008), *CV-ANOVA for Significance Testing of PLS and OPLS Models*, Journal of Chemometrics, 22, 594-600.

Eriksson, L., Trygg, J., and Wold, S., (2009), *PLS-Trees: A Top-Down Clustering Approach*, Journal of Chemometrics, 23, 569-580.

Eriksson, L., and Wold, S., (2010), *A Graphical Index of Separation (GIOS) in Multivariate Modeling*, Journal of Chemometrics, 24, 779-789.

Eriksson, L., Rosén, J., Johansson, E., Trygg, J., (2012), *Orthogonal PLS (OPLS) Modeling for Improved Analysis and Interpretation in Drug Design*, Molecular Informatics, 31, 414-419.

Esbensen, K., and Geladi, P., (1989), *Strategy of Multivariate Image Analysis (MIA)*, Chemometrics and Intelligent Laboratory Systems, 7, 67-86.

Espina, J.R., Shockcor, J.P., Herron, W.J., Car, B.D., Contel, N.R., Ciaccio, P.J., Lindon, J.C., Holmes, E., and Nicholson, J.K., (2001), *Detection of In-vivo Biomarkers of Phospholipidosis Using NMR-based Metabonomic Approaches*, Magnetic Resonance in Chemistry, 39, 559–565.

Etemad, K., and Chellappa, R., (2004), *Discriminant Analysis for Recognition of Human Face Images*, Journal of Optical Society of America, 14, 1724-1733.

Fearn, T., Riccioli, C., Garrido-Varo, A., Guerrero-Ginal, J.E., (2009), *On the Geometry of SNV and MSC*, Chemometrics and Intelligent Laboratory Systems, 96, 22-26.

Fisher, R.A., (1936), *The use of Multiple Measurements in Taxonomic Problems*, Ann. Eugenics, 7, 179-188.

Flåten, G.R., and Walmsley, A.D., (2004), *A Design of Experiment Approach Incorporating Layered Designs for Choosing the Right Calibration Model*, Chemometrics and Intelligent Laboratory Systems, 73, 55-66.

Fotakis, C., Christodouleas, D., Kokkotou, K., Zervou, M., Zoumpoulakis, P., Moulos, P., Liouni, M., and Calokerinos, A., (2013), *NMR metabolite profiling of Greek grape marc spirits*, Food Chemistry, 138, 1837-1846.

Frost, K., Johansson, E., Shanks, R., Goto, S., and Kirwan, G.M., (2013), *Thermoplastic starch films: DOE and O2PLS methodology for optimization and increased understanding of polymer processing*, Polymer Testing, 32, 343-352.

Frank, I.E., and Friedman, J.H., (1993), *A Statistical View of Some Chemometrics Regression Tools*, Technometrics, 35, 109-135.

Frank, I., (1995), *Modern Nonlinear Regression Methods*, Chemometrics and Intelligent Laboratory Systems, 27, 1-19.

Fransson, M., and Folestad, S., (2006), Real-time Alignment of Batch Process Data Using COW for Online Process Monitoring, Chemometrics and Intelligent Laboratory Systems, 84, 56-61.

Gabrielsson, J., Lindberg, N.O., and Lundstedt, T., (2002), *Multivariate Methods in Pharmaceutical Applications*, Journal of Chemometrics, 16, 141-160.

Gabrielsson, J., Jonsson, H., Airiau, C., Schmidt, B., Escott, R., and Trygg, J., (2006), *OPLS Methodology for Analysing Pre-processing of Spectroscopic Data*, Journal of Chemometrics, 20, 362-369.

Gallagher, N.B., Shaver, J.M., Martin, E.B., Morris, J., Wise, B.M., Windig, W., (2004), *Curve Resolution for Multivariate Images with Applications to TOF-SIMS and Raman*, Chemometrics and Intelligent Laboratory Systems, 73, 105-117.

Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gordon, E.M., (1994), *Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries*, Journal of Medicinal Chemistry, 37, 1233-1251.

Garcia-Munoz, S., Kourti, T., and MacGregor, J.F., (2010), *Model Predictive Monitoring for Batch Processes*, Industrial & Engineering Chemistry Research, 43, 59-29-5941.

Gauchi, J.P., and Chagnon, P., (2001), *Comparison of Selection Methods of Explanatory Variables in PLS Regression with Application to Manufacturing Process Data*, Chemometrics and Intelligent Laboratory Systems, 58, 171-193.

Geladi, P., MacDougall, D., and Martens, H., (1985), *Linearization and Scatter-correction for Near-infrared Reflectance Spectra of Meat*, Applied Spectroscopy, 3, 491-500.

Geladi, P., Isaksson, H., Lindqvist, L., Wold, S., and Esbensen, K., (1989), *Principal Component Analysis of Multivariate Images*, Chemometrics and Intelligent Laboratory Systems, 5, 209-220.

Geladi, P., (1992), *Some Special Topics in Multivariate Image Analysis*, Chemometrics and Intelligent Laboratory Systems, 14, 375-390.

Gillet, V.J., (2002), *Reactant- and Product-based Approaches to the Design of Combinatorial Libraries*, Journal of Computer-Aided Molecular Design, 16, 371-380.

Giraud, E., Luttmann, C., Lavelle, F., Riou, J.F., Mailliet, P., and Laoui, A., (2000), *Multivariate Data Analysis using D-optimal Designs, Partial Least Squares, and Response Surface Modelling, A Directional Approach for the Analysis of Farnesyltransferase Inhibitors*, Journal of Medicinal Chemistry, 43, 1807-1816.

Goodford, P., (1996), *Multivariate Characterization of Molecules for QSAR Analysis,* Journal of Chemometrics, 10, 107-117.

Gonzalez-Martinez, J.M., Ferrer, A., and Westerhuis, J.A., (2011), *Real-time Synchronization of Batch Trajectories for On-line Multivariate Statistical Process Control using Dynamic Time Warping*, Chemometrics and Intelligent Laboratory Systems, 105, 195-206.

Gordon, E.M., Barret, R.W., Dower, W.J., Fodor, S.P.A., and Gallop, M.A., (1994), *Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions*, Journal of Medicinal Chemistry, 37, 1385-1401.

Gottfries, J., Blennow, K., Wallin, A., and Gottfries, C.G., (1995), *Diagnosis of Dementias Using Partial Least Squares Discriminant Analysis*, Dementia, 6, 83-88.

Gottfries, J., Depui, H., Fransson, H., Jongeneelen, M., Josefsson, M., Langkilde, F.W., and Witte, D.T., (1996), *Vibrational Spectrometry for the Assessment of Active Substance in Metoprolol Tablets: A Comparison Between Transmission and Diffuse Reflectance Near-Infrared Spectrometry*, Journal of Pharmaceutical and Biomedical Analysis, 14, 1495-1503.

Gottfries, J., Sjögren, M., Holmberg, B., Rosengren, L., Davidsson, P., and Blennow, K., (2004), *Proteomics for Drug Target Discovery*, Chemometrics and Intelligent Laboratory Systems, 73, 47-53.

Grainger, D.J., (2003), *Megavariate Statistics Meets High Data-density Analytical Methods: The Future of Medical Diagnostics*, IRTL Reviews 1, 1-6.

Granberg, R., (1998), *Solubility and Crystal Growth of Paracetamol in Various Solvents,* Ph.D. Thesis, Royal Institute of Technology, Stockholm, Sweden.

Güler, C., (2007), Characterization of Turkish Bottled Waters using Pattern Recognition Methods, Chemometrics and Intelligent Laboratory Systems, 86, 86-95.

Gunnarsson, I., Andersson, P., Wikberg, J., and Lundstedt, T., (2003), *Multivariate Analysis of G Protein-coupled Receptors*, Journal of Chemometrics, 17, 82-92.

Gustafsson, A., (1993), *QFD and Conjoint Analysis – The Key to Customer Oriented Products*, Ph.D. Thesis, Linköping University, Linköping, Sweden.

Halstensen, M., Kaku-Arvoh, B., Amundsen, L., and Hoffmann, R., (2013), *Online estimation of wax depositions thickness in single-phase sub-sea pipelines based on acoustic chemometrics: A feasibility study*, Fuel, 105, 718-727.

Hammett, L.P., (1970), *Physical Organic Chemistry*, 2nd edn., McGraw-Hill, New York.

Hanafi, M., Kohler, A., and Quannari, E.M, (2010), *Shedding New Ligths on Hierarchical Principal Component Analysis*, Journal of Chemometrics, 24, 703–709.

Hand, D.J., (1998), *Data Mining: Statistics and More?*, American Statistician, 52, 112-188.

Hansch, C., and Leo, A.J., (1970), *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York.

Hartnett, M.K., Lightbody, G., and Irwin, G.W., (1999), *Identification of State Models Using Principal Components Analysis*, Chemometrics and Intelligent Laboratory Systems, 46, 181-196.

Hassani, S., Martens, H., Qannari, E.M., Hanafi, M., and Kohler, A., (2012), *Model validation and error estimation in multi-block partial least squares regression*, Chemometrics and Intelligent Laboratory System, 117, 42-53.

Hastie, T., Tibshirani, R., and Walther, G., (2000), *Estimating the Number of Data Clusters via the Gap Statistic*, www.stanford.edu/~tibs/ftp/gap.ps.

Hellberg, S., (1986), *A Multivariate Approach to QSAR*, Ph.D. Thesis, Umeå University, Umeå, Sweden.

Hellberg, S., Sjöström, M., and Wold, S., (1986), *The Prediction of Bradykinin Potentiating Potency of Pentapeptides. An Example of a Peptide Quantitative Structure-Activity Relationship*, Acta Chemica Scandinavica, B40, 135-140.

Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S., and Andrews, P., (1991), *Minimum Analogue Peptide Sets (MAPS) for Quantitative Structure-Activity Relationships,* International Journal of Peptide and Protein Research, 37, 414-424.

Hermens, J.L.M., (1989), *Quantitative Structure-Activity Relationships of Environmental Pollutants.* In: Hutzinger, O., (Ed.), Handbook of Environmental Chemistry, Vol 2E, Reactions and Processes. Springer-Verlag, Berlin, 1989, pp. 111-162.

Hunter, J.S., (1986), *The Exponentially Weighted Moving Average*, Journal of Quality Technology, 18, 203-210.

Höskuldsson, A., (1996), *Prediction Methods in Science and Technology*, Thor Publishing, Copenhagen, Denmark.

Höskuldsson, A., (1998), *The Heisenberg Modelling Procedure and Application to Nonlinear Modelling*, Chemometrics and Intelligent Laboratory Systems, 44, 15-30.

Höskuldsson, A., (2001), *Causal and Path Modelling*, Chemometrics and Intelligent Laboratory Systems, 58, 287-311.

Indahl, U.G., and Naes, T., (1998), *Evaluation of Alternative Spectral Feature Extraction Methods of Textural Images for Multivariate Modeling*, Journal of Chemometrics, 12, 261-278.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, John Wiley, New York. ISBN 0-471-62267-2.

Janné, K., Pettersen, J., Lindberg, N.O., and Lundstedt, T., (2001), *Hierarchical Principal Component Analysis (PCA) and Projection to Latent Structure Technique (PLS) on Spectroscopic Data as a Data Pretreatment for Calibration*, Journal of Chemometrics, 15, 203-213.

Jellum, E., Björnsson, I., Nesbakken, R., Johansson, E., and Wold, S., (1981), *Classification of Human Cancer Cells by means of Capillary Gas Chromatography and Pattern Recognition Analysis*, Journal of Chromatography, 217, 231-237.

Johansson, D, and Lindgren, P., (2002), *Analysis of Microarray Data Using a Multivariate Approach*, Masters Thesis in Bioinformatics, Umeå University, Umeå, Sweden.

Johansson, D., Lindgren, P., and Berglund, A., (2003), *A Multivariate Approach Applied to Microarray Data for Identification of Genes with Cell Cycle-Coupled Transcription*, Bioinformatics, 19, 467-473.

Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., (1989a), *Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids*, Quantitative Structure-Activity Relationships, 8, 204-209.

Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., (1989b), *A Multivariate Approach to Saccharide Quantitative Structure-Activity Relationships Exemplified by Two Series of 9-Hydroxyellipticine Glycosides*, Acta Chemica Scandinavica, 43, 286-289.

Jonsson, J., Eriksson, L., Hellberg, S., Lindgren, F., Sjöström, M., and Wold, S., (1991), *A Multivariate Representation and Analysis of DNA Sequence Data*, Acta Chemica Scandinavica, 45, 186-192.

Jonsson, J., (1992), *Quantitative Sequence-Activity Modelling (QSAM),* Ph. D. Thesis, Umeå University, Umeå, Sweden.

Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S., (1993), *Quantitative Sequence-Activity Models (QSAM) – Tools for Sequence Design*, Nucleic Acids Research, 21, 733-739.

Jonsson, P., Sjöström, M., Wallbäcks, L., and Antti, H., (2004), *Strategies for Implementation and Validation of On-line Models for Multivariate Monitoring and Control of Wood Chips Properties*, Journal of Chemometrics, 18, 203-207.

Jonsson, P., Sjövik-Johansson, E., Woulikainen, A., Lindberg, J., Schuppe-Koistinen, I., Kusano, M., Sjöström, M., Trygg, J., Moritz, T., and Antti, H., (2006), *Predictive Metabolite Profiling Applying Hierarchical Multivariate Curve Resolution to GC-MS Data - A Potential Tool for Multi-Parametric Diagnosis*, Journal of Proteome Research, 5,1407-1414.

Jos, A., Moreno, I., Gonzalez, A.G., Repetto, G., and Camean, A.M., (2004), *Differentiation of Sparkling Wines (Cava and Champagne) According to their Mineral Content*, Talanta, 63, 377-382.

Kassidas, A., MacGregor, J.F., and Taylor, P.A., (1998), *Synchronization of Batch Trajectories Using Dynamic Time Warping*, AlChE Journal, 44, 864-875.

Kawabe, T., Tomitsuka, T., Kajiro, T., Kishis, N., and Toyooka, T., (2013), *Ternary isocratic mobile phase optimization utilizing resolution design space based on retention time and peak width modeling*, Journal of Chromatography A, 1273, 95-104.

Kettaneh-Wold, N., MacGregor, J.F., Dayal, B., and Wold, S., (1994), *Multivariate Design of Process Experiments (M-DOPE)*, Chemometrics and Intelligent Laboratory Systems, 23, 39-50.

Kettaneh, N., Berglund, S., and Wold, S., (2005), *PCA and PLS with Very Large Data Sets*, Computational Statistics & Data Analysis, 48, 69-85.

Kim, K.H., (1993), *Non-linear Dependence in Comparative Molecular Field Analysis*, Journal of Computer-Aided Molecular Design, 7, 71-82.

Kim, S., Kano, M., Nakagawa, H., and Hasebe, S., (2011), *Estimation of Active Pharmaceutical Ingredients Content using Locally Weighted Partial Least Squares and Statistical Wavelength Selection*, International Journal of Pharmaceutics, 421, 269-274.

Kimura, T., Miyashita, Y., Funatsu, K., and Sasaki, S., (1996), *Quantitative Structure-Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression*, Journal of Chemical Information and Computer Science, 36,185-189.

Kjaer Pedersen, D., Martens, H., Pram Nielsen, J., Engelsen, S.B., (2002), *Near-infrared Absorption and Scattering Separated by Extended Inverted Signal Correction (EISC): Analysis of Near-infrared Transmittance Spectra of Single Wheat Seeds*, Applied Spectroscopy, 56, 1206-1214.

Kohonen, J., Reinikainen, S.P., Aaljoki, K., and Höskuldsson, A., (2009), *Non-linear PLS Approach in Score Surface*, Chemometrics and Intelligent Laboratory Systems, 97, 159-163.

Kourti, T., (2002), *Process Analysis and Abnormal Situation Detection: From Theory to Practice*, IEEE Control Systems Magazine, October 2002, 10-25.

Kourti, T., (2003), *Multivariate Dynamic Data Modeling for Analysis and Statistical Process Control of Batch Processes*, *Start-ups and Grade Transitions*, Journal of Chemometrics, 17, 93-109.

Kourti, T., (2005), *Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry*, International Journal of Adaptive Control and Signal Processing, 19, 213-246.

Kriegl, J.M., Eriksson, L., Arnhold, T., Beck, B., Johansson, E., and Fox, T., (2005), *Multivariate Modeling of Cytochrome P450 3A4 Inhibition*, European Journal of Pharmaceutical Sciences, 24, 451-463.

Kristal, B.S., (2002), *Practical Considerations and Approaches for Entry-level Megavariate Analysis*. http://mickey.utmem.edu/papers/bioinformatics_02/pdfs/Kristal.pdf. Accessed 2005-11-14.

Kubinyi, H., (1990), *Quantitative Structure-Activity Relationships and Molecular Modelling in Cancer Research*, Journal Cancer Research & Clinical Oncology, 116, 529-537, 1990.

Kvalheim, O.M., (1989), *Interpretation of Latent-Variable Regression models*, Chemometrics and Intelligent Laboratory Systems, 7, 39-51.

Könemann, H., (1981), *Quantitative Structure-Activity Relationships in Fish Studies. Part 1: Relationship for 50 Industrial Pollutants*, Toxicology, 19, 209-221.

Larsson, U., Carlson, R., and Leroy. J., (1993), *Synthesis of Amino Acids with Modified Principal Properties. 1. Amino Acids with Fluorinated Side Chains*, Acta Chemica Scandinavica, 47, 380-390.

Lauri, D., Martinez, M., Salcedo, J.V., and Sanchis, J., (2010), *PLS-based Model Predictive Control Relevant Identification: PLS-PH Algorithm*, Chemometrics and Intelligent Laboratory Systems, 100, 118-126.

Laursen, K., Söndergaard Frederiksen, S., Leuenhagen, C., and Bro, R., (2010), *Chemometric Quality Control of Chromatographic Purity*, Journal of Chromatography A, 1217, 6503-6510.

Leardi, R., (2001), *Genetic Algorithms in Chemistry and Chemometrics: A Review*, Journal of Chemometrics, 15, 559-569.

Lee, H.W., Lee, M.W., and Park, J.M., (2009), *Multiscale Extension of PLS Algorithm for Advanced Online Process Monitoring*, Chemometrics and Intelligent Laboratory Systems, 98, 201-212.

Lestander, T., Holmberg, C., Stenberg, L., and Lehtonen, R., (2012), *Towards multivariate statistical process control in the wood pellet industry*, Biomass and Bioenergy, 45, 152-158.

Lewi, P.J., (1995), *Spectral Mapping of Drug-Test Specificities*, In: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim.

Li, H., Liang, Y., and Xy, Q., (2009), *Support Vector Machine and its Applications in Chemistry*, Chemometrics and Intelligent Laboratory Systems, 95, 188-198.

Liao, T.W., (2005), *Clustering of Time Series Data – A Survey*, Pattern Recognition, 38, 1857-1874.

Liebmann, B., Friedl, A., and Varmuza, K., (2010), *Applicability of Near-infrared Spectroscopy for Process Monitoring in Bioethanol Production*, Biochemical Engineering Journal, 52, 187–193.

Lied, T.T., Geladi, P., and Esbensen, K.H., (2000), *Multivariate Image Regression (MIR): Implementation of Image PLSR – First Forays*, Journal of Chemometrics, 14, 585-598.

Lindberg, N.O., and Gabrielsson, J., (2004), *Use of Software to Facilitate Pharmaceutical Formulation – Experiences from a Tablet Formulation*, Journal of Chemometrics, 18, 133-138.

Linderholm, J., and Lundberg, E., (1994), *Chemical Characterization of Various Archaeological Soil Samples using Main and Trace Elements Determined by Inductively Coupled Plasma Atomic Emission Spectrometry,* Journal of Archaeological Science, 21, 303-314.

Lindgren, F., and Geladi, P., (1992), *Multivariate Spectrometric Image Analysis – An Illustrative Study with two Constructed Examples of Metal Ions in Solution*, Chemometrics and Intelligent Laboratory System, 14, 397-412.

Lindgren, F., Geladi, P., Rännar, S., and Wold, S., (1994), *Interactive Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms*, Journal of Chemometrics 8, 349-363.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., and Eriksson, L., (1996), *Model Validation by Permutation Tests*, Journal of Chemometrics, 10, 521-532.

Lindgren, Å., Sjöström, M., and Wold, S., (1996), *Quantitative Structure-Effect Relationships for Some Technical Nonionic Surfactants*, JAOCS, 7, 863-875.

Lindgren, Å., (2000), *Use of Multivariate Methods and DOE to Improve Industrial-Scale Production Quality of a Cellulose Derivative*. Journal of Chemometrics, 14, 657-665.

Lindström, A., Pettersson, F., and Linusson, A., (2009), *Quantitative Protein Descriptors for Secondary Structure Characterization and Protein Classification*, Chemometrics and Intelligent Laboratory Systems, 95, 74-85.

Linusson, A., Gottfries, J., Lindgren, F., and Wold, S., (2000), *Statistical Molecular Design of Building Blocks for Combinatorial Chemistry*, Journal of Medicinal Chemistry, 43, 1320-1328.

Linusson, A., (2000), *Efficient Library Selection in Combinatorial Chemistry*, Ph.D. Thesis, Umeå University, Umeå, Sweden.

Linusson, A., Gottfries, J., Olsson, T., Örnskov, E., Folestad, S., Nordén, B., and Wold, S., (2001), *Statistical Molecular Design, Parallel Synthesis, and Biological Evaluation of a Library of Thrombin Inhibitors*, Journal of Medicinal Chemistry, 44, 3424-3439.

Lockhart, D.J., and Winzeler, E.A., (2000), *Genomics, Gene Expression and DNA Arrays*, Nature, 405, 827-836.

Long, I., Andersson, P., Siefert, E., and Lundstedt, T., (2004), *Multivariate Analysis of Five GPCR Receptor Classes*, Chemometrics and Intelligent Laboratory Systems, 73, 95-104.

Louwerse, D.J., Tates, A.A., Smilde, A.K., Koot, G.L.M., and Berndt, H., (1999), *PLS Discriminant Analysis with Contribution Plots to Determine Differences Between Parallel Batch Reactors in the Process Industry*, Chemometrics and Intelligent Laboratory Systems, 46, 197-206.

Lowe, G., (1995), *Combinatorial Chemistry*, Chemical Society Reviews, 24, 309-317.

Lundstedt, T., (1991), *A QSAR Strategy for Screening of Drugs and Predicting Their Clinical Activity*, Drug News & Perspectives, 4, 468-475.

Lundstedt, T., Carlson, R., and Shabana, R., (1987), *Optimum Conditions, for the Willgerodt-Kindler Reaction. 3. Amine Variation*, Acta Chemica Scandinavica, B41, 157-163.

Lundstedt, T., Clementi, S., Cruciani, G., Pastor, M., Kettaneh, N., Andersson, P.M., Linusson, A., Sjöström, M., Wold, S., and Nordén. B., (1997), *Intelligent Combinatorial Libraries*, In: H. van de Waterbeemd, B. Testa and G. Folkers, eds., Computer-Assisted Lead Finding and Optimization, Current Tools for Medicinal Chemistry, Wiley-VCH, Weinheim.

Luukkonen, P., Fransson, M., Niklasson Björn, I., Hautala, J., Lagerholm, B., and Folestad, S., *Real-Time Assessment of Granule and Tablet Properties Using In-line Data from a High-shear Granulation Process*, (2008), Journal of Pharmaceutical Sciences, 97, 938-947.

Löfstedt, T., and Trygg, J., (2011), *OnPLS – a Novel Multiblock Method for the Modelling of Predictive and Orthogonal Variation*, Journal of Chemometrics, 25, 441-455.

Löfstedt, T., Eriksson, K., Wormbs, G., and Trygg, J., (2012a), *Bi-modal OnPLS*, Journal of Chemometrics, 26, 236-245.

Löfstedt, T., Hanafi, M., Mazerolles, G., and Trygg, J., (2012b), *OnPLS Path Modelling*, Chemometrics and Intelligent Laboratory Systems, 118, 139-149.

Macedo, I.S.M., Sousa-Gallagher, M.J., Oliveira, J.C., and Byrne, E.P., (2013), *Quality by design for packaging of granola breakfast product*, Food Control, 29, 438-443.

MacFie, H., Moore, P.B., and Wakeling, I., (1999), *Changes in the Sensory Properties and Consumer Preferences for Dessert Apples*, Apples and Pears Research Council, UK.

MacGregor, J.F., and Nomikos, P., (1992), *Monitoring Batch Processes*, Proceedings NATO Advanced Study Institute for Batch Processing Systems Eng., May 29 – June 7, 1992, Antalya, Turkey.

MacGregor, J.F., and Kourti, T., (1995), *Statistical Process Control of Multivariate Processes*, Control Eng. Practice, 3, 403-414.

MacGregor, J.F., (1996), *Using On-Line Process Data to Improve Quality*, ASQC Statistics Division Newsletter, 16, 6-13.

Maitra, R., (2001), *Clustering Massive Data Sets with Applications in Software Metrics and Tomography*, Technometrics, 43, 336-346.

Mallat, S., (1998), *A Wavelet Tour of Signal Processing*, Academic Press

Martens, H., and Naes, T., (1989), *Multivariate Calibration*, John Wiley, New York.

Martens, H., and Martens, M., (2000), *Modified Jack-Knife Estimation of Parameter Uncertainty in Bilinear Modeling (PLSR)*, Food Quality and Preference, 11, 5-16.

Martens, H., and Martens, M., (2001), *Significance Testing with Cross-validation versus ANOVA*, in Multivariate Analysis of Quality: An Introduction, Harald Martens, M. Martens, ISBN: 978-0-471-97428-4.

Martens, H., Pram Nielsen, J., and Engelsen, S.B., (2003), *Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures*, Analytical Chemistry, 75, 394-404.

Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H., (1995), *Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery*, Journal of Medicinal Chemistry, 38, 1431-1436.

Marvanova, S., Nagata, Y., Wimmerova, M., Sykorova, J., Hynkova, K., and Damborsky, J., (2001), *Biochemical Characterization of Broad-specificity Enzymes Using Multivariate Experimental Design and a Colorimetric Microplate Assay: Characterization of the Haloalkane Dehalogenase Mutants*, Journal of Microbiological Methods, 44, 149-157.

Massart, B., (1997), *Environmental Monitoring and Forecasting by Means of Multivariate Methods*, Ph.D. Thesis, University of Bergen, Bergen, Norway.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., and Kaufman, L., (1988), *Chemometrics: A Textbook*, Elsevier.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., and Smeyers-Verbeke, J., (1998), *Handbook of Chemometrics and Qualimetrics*, Elsevier.

McCloskey, J.T., Newman, M.C., and Clark, S.B., (1996), *Predicting the Relative Toxicity of Metal Ions Using Ion Characteristics: Microtox Bioluminescence Assay*, Environmental Toxicology and Chemistry, 15, 1730-1737.

McCready, C., (2011), Model Predictive Multivariate Control, In: Proc. 2nd European Conference on Process Analytics and Control Technology, pp. 82-86.

McEwan, J.A., Earthy, P.J., and Ducher, C., (1998), *Preference Mapping – A Review*, Review No. 6, Campden & Chorleywood Food Research Association, Gloucestershire, UK.

McEwan, J.A., and Ducher, C., (1998), *Preference Mapping – Case Studies*, Review No. 7, Campden & Chorleywood Food Research Association, Gloucestershire, UK.

Michailidis, G., and de Leeuw, J., (1998), *The GIFI System of Descriptive Multivariate Analysis*, Statistical Science, 13, 307-336.

Michielan, L., and Moro, S., (2010), *Pharmaceutical Perspectives of Nonlinear QSAR Strategies*, Journal of Chemical Information and Modelling, 50, 961-978.

Mirschel, G., Savchuk, O., Scherzer, T., and Genest, B., (2013), *Process control of printing processes with in-line NIR spectroscopy and elimination of the influence of the substrate on the prediction of the coating weight*, Progress in Organic Coatings, 76, 86-93.

Moberg, L., Karlberg, B., Blomqvist, S., and Larsson, U., (2000), *Comparison Between a New Application of Multivariate Regression and Current Spectroscopy Methods for the Determination of Chlorophylls and Their Corresponding Pheopigments*, Analytica Chimica Acta, 411, 137-143.

Moberg, L., and Karlberg, B., (2001), *Validation of a Multivariate Calibration Method for the Determination of Chlorophyll a, b, and c and Their Corresponding Pheopigments*, Analytica Chimica Acta, 450, 143-153.

Mokgalaka-Matlala, N.S., Regnier, T., Combrinck, S., Kouekam, C.R., and Weiersbye, I.M., (2013), *Chemometrics and vibrational spectroscopy as green tools for mine phytoremediation strategies*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 100, 138-143.

Mullet, G.M., (1976), *Why Regression Coefficients have the Wrong Sign*, Journal of Quality Technology, 8, 121-126.

Munck, L., Nörgaard, L., Engelsen, S.B., Bro, R., and Andersen, C., (1998), *Chemometrics in Food Science – A Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance*, Chemometrics and Intelligent Laboratory, 44, 31-60.

Musumarra, G., Barresi, V., Condorelli, D.F., Fortuna, C.G., and Scirè, S., (2004) *Potentialities of Multivariate Approaches in Genome-based Cancer Research: Identification of Candidate Genes for New Diagnostics by PLS Discriminant Analysis*, Journal of Chemometrics, 18, 125-132.

Naes, T., and Indahl, U., (1998), *A Unified Description of Classical Classification Methods for Multicollinear Data,* Journal of Chemometrics, 12, 205-220.

Naes, T., and Mevik, B.H., (1999), *The Flexibility of Fuzzy Clustering Illustrated by Examples*, Journal of Chemometrics, 13, 435-444.

Naes, T., Isaksson, T., Fearn, T., and Davies, T., (2002), *A User-friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, UK. ISBN: 0-95286662-5.

Nakagawa, H., Tajima, T., Kano, M., Kim, S., Hasebe, S., Suzuki, T., and Nakagami, H., (2012), *Evaluation of Infrared Reflection Absorption Spectroscopy Measurements and Locally Weighted Partial Least Squares for Rapid Analysis of Residual Drug Substances in Cleaning Processes*, Analytical Chemistry, 84, 3820-3826.

Nicholson, J.K., Lindon, J.C., Holmes, E., (1999), *Metabonomics: Understanding the Metabolic Responses of Living Systems to Pathophysiological Stimuli via Multivariate Statistical Analysis of Biological NMR Spectroscopic Data*, Xenobiotica, 29, 1181-1189.

Nicholson, J.K, Connelly, J., Lindon, J.C., and Holmes, E., (2002), *Metabonomics: a Platform for Studying Drug Toxicity and Gene Function*, Nature Reviews, 1, 153-162.

Nijhuis, A., de Jong, S., and Vandeginste, B.G.M., (1997), *Multivariate Statistical Process Control in Chromatography*, Chemometrics and Intelligent Laboratory Systems, 38, 51-62.

Nilsson, D., (2005), *Prediction of Wood Species and Pulp Brightness from Roundwood Measurements*, PhD Thesis, Umeå University, Umeå, Sweden.

Nomikos, P., and MacGregor, J.F., (1995a), *Multivariate SPC Charts for Monitoring Batch Processes,* Technometrics, 37, 41-59.

Nomikos, P., and MacGregor, J.F., (1995b), *Multiway Partial Least Squares in Monitoring of Batch Processes*, Chemometrics and Intelligent Laboratory System, 30, 97-108.

Nomizu, M., Iwaki, T., Yamashita, T., Inagaki, Y., Asano, K., Akamatsu, M., and Fujita, T., (1993), *Quantitative Structure-Activity Relationship (QSAR) Study of Elastase Substrates and Inhibitors*, International Journal of Peptide and Protein Research, 42, 216-226.

Nordahl, Å., and Carlson, R., (1993), *Exploring Organic Synthetic Procedures*, Topics in Current Chemistry, 166, 1-64.

Norinder, U., (1996), *Single and Domain Mode Variable Selection in 3D QSAR Applications*, Journal of Chemometrics, 10, 95-105.

Norinder, U., Sjöberg, P., and Österberg, T., (1998), *Theoretical Calculation and Prediction of Brain-Blood Partitioning or Organic Solutes using Molsurf Parametrization and PLS Statistics*, Journal of Pharmaceutical Sciences, 87, 952-959.

Norinder, U., (2003), *Support Vector Machine Models in Drug Design: Applications to Drug Transport Processes and QSAR Using Simplex Optimisations and Variable Selection*, Neurocomputing, 55, 337-346.

Nouwen, J., Lindgren, F., Hansen, B., Karcher, W., Verhaar, H.J.M., and Hermens, J.L.M., (1997), *Classification of Environmentally Occurring Chemicals Using Structural Fragments and PLS Discriminant Analysis,* Environmental Science and Technology, 31, 2313-2318.

Nyström, Å., Andersson, P.M., and Lundstedt, T., (2000), *Multivariate Data Analysis of Topographically Modified a-Melanotropin Analogues Using Auto- and Cross-Auto Covariances*, Quantitative Structure-Activity Relationships, 19, 264-269.

Oliveri, P., Di Egidio, V., Woodcock, I., and Downey, G., (2011), *Application of class-Modelling Techniques to Near Infrared Data for Food Authentication Purposes*, Food Chemistry, 125, 1450–1456.

Olsson, I.M., Gottfries, J., and Wold, S., (2004a), *D-optimal Onion Design in Statistical Molecular Design*, Chemometrics and Intelligent Laboratory Systems, 73, 37-46.

Olsson, I.M., Gottfries, J, and Wold, S., (2004b), *Controlling Coverage of D-optimal Onion Designs and Selections*, Journal of Chemometrics, 18, 548-557.

Oprea, T.I., (2000), *Property Distribution of Drug-Related Chemical Data-Bases*, Journal of Computer-Aided Molecular Design, 14, 251-264.

Oprea, T.I., and Gottfries, J., (2001), *Chemography: The Art of Navigating in Chemical Space*, Journal of Combinatorial Chemistry, 3, 157-166.

Peres-Neto, P.R., Jackson, D.A., and Somers, K.M., (2005), *How Many Principal Components? Stopping Rules for Determining the Number of Non-trivial Axes Revisited*, Computational Statistics & Data Analysis, 49, 974-997.

Peterson, J., (2010), *What Tour ICH Q8 Design Space Needs: A Multivariate Predictive Distribution*, Pharma Manufacturing, 9, 1-12.

Phatak, A., and DeJong, S., (1997), *The Geometry of PLS*, Journal of Chemometrics, 11, 311-338.

Piazza, L., Gigli, J., and Ballabio, D., (2007), *On the Application of Chemometrics for the Study of Acoustic Mechanical Properties of Crispy Bakery Products*, Chemometrics and Intelligent Laboratory Systems, 86, 52-59.

Prats-Montalban, J.M., Ferrer, A., Malo, J.L., and Gorbena, J., (2006), *A comparison of Different Discriminant Analysis Techniques in a Steel Industry Welding Process*, Chemometrics and Intelligent Laboratory Systems, 80, 109-119.

Qin, S.J., and McAvoy, T.J., (1992), *Non-Linear PLS Modelling Using Neural Networks*, Computation and Chemical Engineering, 16, 379-391.

Quintas, G., Portillo, N., Garcia-Canaveras, J.C., Castell, J.V., Ferrer, F., and Lahoz, A., (2012), *Chemometric Approaches to Improve PLS-DA Model Outcome for Predicting Human Non-alcoholic Fatty Liver Disease using UPLC-MS as a Metabolic Profiling Tool*, Metabolomics, 8, 86-98.

Ramos, E.U., Vaes, W.H.J., Verhaar, H.J.M., and Hermens, J.L.M., (1997), *Polar Narcosis: Designing a Suitable Training Set for QSAR Studies*, Environmental Science & Pollution Research, 4, 83-90.

Rius, A., Ruisanchez, I., Callao, M.P., Rius, F.X., (1998), *Reliability of Analytical Systems: Use of Control Charts, Time Series Models and Recurrent Neural Networks*, Chemometrics and Intelligent Laboratory Systems, 40, 1-18.

Rodrigez-Pulido, F.J., Barbin, D.F., Sun, D.W., Gordillo, B., Gonzalez-Miret, M.L., and Heredia, F.J., (2013), *Grape seed characterization by NIR hyperspecral imaging*, Postharvest Biology and Technology, 76, 74-82.

Rozet, E., Lebrun, P., Debrus, B., Boulanger, B., and Hubert, P., (2013), *Design spaces for analytical methods*, Trends in Analytical Chemistry, 42, 157-167.

Rudnitskaya, A., Delgadillo, I., Legin, A., Rocha, S.M., Costa, A.M., and Simoes, T., (2007), *Prediction of the Port Wine Age using an Electronic Tongue*, Chemometrics and Intelligent Laboratory Systems, 87, 50-56.

Rücker, C., Rücker, G., and Meringer, M., (2007), Y-*Randomization and its Variants in QSPR/QSAR*, Journal of Chemical Information and Modeling, 47, 2345-2357.

Rännar, S., MacGregor, J.F., and Wold, S., (1998), *Adaptive Batch Monitoring Using Hierarchical PCA*, Chemometrics and Intelligent Laboratory Systems, 41, 73-81.

Rännar, S., and Andersson, P.L., (2010), *A Novel Approach using Hierarchical Clustering to Select Industrial Chemicals for Environmental Impact Assessment*, Journal of Chemical Information and Modeling, 50, 30-36.

Sandasi, M., Kamatou, G.P.P. and Viljoen, A.M., (2012), *An untargeted metabolomic approach in the chemotaxonomic assessment of two Salvia species as a potential source of alfa-bisabolol*, Phytochemistry, 84, 94-101.

Sandberg, M., Sjöström, M., and Jonsson, J., (1996), *A Multivariate Characterization of tRNA Nucleosides,* Journal of Chemometrics, 10, 493-508.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., (1998), *New Chemical Dimensions Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids*, Journal of Medicinal Chemistry, 41, 2481-2491.

Savitzky, A., and Golay, M.J.E., (1964), *Smoothing and Differentiation by Simplified Least Squares Procedures*, Analytical Chemistry, 36, 1627-1632.

Schaper, K.J., (1991), *QSAR Analysis of Chiral Drugs Including Stereoisomer Combinations*, In: Silipo, C., and Vittoria, A, (Eds.), QSAR: Rational Approaches to the Design of Bioactive Compounds, Elsevier Science Publishers, Amsterdam, 1991, pp. 25-32.

Schüürmann, G., Rayasamuda, K., and Kristen, U., (1996), *Structure-Activity Relationships for Chloro- and Nitrophenol Toxicity in the Pollen Tube Growth Test*, Environmental Toxicology and Chemistry, 15, 1702-1708.

Searson, D., Willis, M., and Montague, G., (2007), *Co-evolution of Nonlinear PLS Model Components*, Journal of Chemometrics, 21, 592-603.

Sekulic, S., Seasholtz, M.B., Wang, Z., Kowalski, B., Lee, S.E., and Holt, B.R, (1993), *Non-linear Multivariate Calibration Methods in Analytical Chemistry*, Analytical Chemistry, 65, 835-845.

Seydel, J.K., Wiese, M., Cordes, H.P., Chi, H.L., Schaper, K.-J., Coats, E.A., Kunz, B., Engel, J., Kutscher, B., and Emig, H., (1991), *QSAR and Modelling of Enzyme Inhibitors, Anticonvulsants and Amphiphilic Drugs Interacting with Membranes*, In:, Silipo, C., and Vittoria, A, (Eds.), QSAR: Rational Approaches to the Design of Bioactive Compounds, Elsevier Science Publishers, Amsterdam, 1991, pp. 367-376.

Shao, J., (1993), Linear Model Selection by Cross-Validation, Journal of the American Statistical Association, 88, 486-494.

Shewhart, W., (1931), *Economic Control of Quality of Manufactured Product*, Van Nostrand, Princeton, N.J.

Singhal, A., and Seborg, D.E., (2002), *Pattern Matching in Historical Batch Data Using PCA*, IEEE Control Systems Magazine, October 2002, 53-63.

Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H., and Wold, S., (1998), *An Evaluation of Orthogonal Signal Correction Applied to Calibration Transfer of Near Infrared Spectra*, Chemometrics and Intelligent Laboratory Systems, 44, 229-244.

Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å., and Martens, H., (1983), *A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables*, Analytica Chimica Acta, 150, 61-70.

Sjöström, M., Wold, S., and Söderström, B., (1986), *PLS Discriminant Plots,* Proceedings of PARC in Practice, Amsterdam, June 19-21, 1985. Elsevier Science Publishers B.V., North-Holland.

Sjöström, M., Eriksson, L., Hellberg, S., Jonsson, J., Skagerberg, B., and Wold, S., (1989), *Peptide QSARs: PLS Modelling and Design in Principal Properties.* In: J.L. Fauchère (ed.): QSAR - Quantitative Structure-Activity Relationships in Drug Design. Proc. 7th European Symposium on QSAR, Sept. 1988, Interlaken, Switzerland. Alan R. Liss, Inc., New York, pp. 131-134.

Sjöström, M., Rännar, S., and Rilfors, L., (1995), *Polypetide Sequence Property Relationships in Escherichica Coli Based on Auto Cross Covariances*, Chemometrics and Intelligent Laboratory Systems, 29, 295-305.

Sjöström, M., Lindgren, Å., and Uppgård, L.L., (1997), *Joint Multivariate Quantitative Structure-Property and Structure-Activity Relationships for a Series of Technical Nonionic Surfactants*, In: F. Chen & G. Schüürmann (eds.), Quantitative Structure-Activity Relationships in Environmental Sciences – VII. Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, June 24-28, 1996, Elsinore, Denmark. SETAC Press, Pensacola, Florida, 1997, pp. 435-449.

Skagerberg, B., Sjöström, M., and Wold, S., (1987), *Multivariate Characterization of Amino Acids by Reversed Phase High Pressure Liquid Chromatography*, Quantitative Structure-Activity Relationships, 6, 158-164.

Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G., and Ebert, C., (1989), *Principal Properties for Aromatic Substituents. A Multivariate Approach for Design in QSAR*, Quantitative Structure-Activity Relationships, 8, 32-38.

Skoglund, A., Brundin, A., and Mandenius C.F., (2004), *Monitoring a Paperboard Machine Using Multivariate Statistical Process Control*, Chemometrics and Intelligent Laboratory Systems, 73, 3-6.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., (1998), *Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization*, Molecular Biology of the Cell, 9, 3273-3297.

Stefanov, Z.I., and Hoo, K.A., (2003), *Hierarchical Multivariate Analysis of Cockle Phenomena*, Journal of Chemometrics, 17, 550-568.

Stenlund, H., Madsen, R., Vivi, A., Calderisi, M., Lundstedt, T., Tassini, M., Carmellini, M., and Trygg, J., (2009), *Monitoring Kidney-Transplant Patients using Metabolomics and Dynamic Modelling*, Chemometrics and Intelligent Laboratory Systems, 98, 45-50.

Stork, C.L., and Kowalski, B.R., (1999), *Distinguishing Between Process Upsets and Sensor Malfunctions Using Sensor Redundancy*, Chemometrics and Intelligent Laboratory Systems, 46, 117-131.

Streefland, M., van Herpen, P.F.G., van de Waterbeemd, B., van der Pol, L.A., Beuvery, E.C., Tramper, J., Martens, D.E., and Toft, M., (2009), *A Practical Approach for Exploration and Modeling of the Design Space of a Bacterial Vaccine Cultivation Process*, Biotechnology and Bioengineering, 104, 492-504.

Strouf, O., (1986), *Chemical Pattern Recognition*, John Wiley, New York.

Ståhle, L., and Wold, S., (1987), *Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study*, Journal of Chemometrics, 1, 185-196.

Suzuki, M., Kusano, M., Takahaski, H., Nakamura, Y., Hayashi, N., Kobayashi, M., Ichikawa, T., Matsui, M., Hirochika, H., Saito, K., (2012), *Rice-Arabidopsis FOX-line Screening with FT-NIR-based Fingerprinting for GC-TOF/MS-based Metabolite Profiling, Metabolomics*, DOI 10.1007/s11306-009-0182-2.

Svensson, O., Josefsson, M., and Langkilde, F.W., (1997), *Classification of Chemically Modified Celluloses Using a Near-Infrared Spectrometer and Soft Independent Modelling of Class Analogy*, Applied Spectroscopy, 51, 1826-1835.

Svensson, O., Josefsson, M., and Langkilde, F.W., (1999), *Reaction Monitoring Using Raman Spectroscopy and Chemometrics*, Chemometrics and Intelligent Laboratory Systems, 49, 49-66.

Svensson, O., Kourti, T., and MacGregor, J.F., (2002), *An Investigation of Orthogonal Signal Correction Algorithms and their Characteristics*, Journal of Chemometrics, 16, 176-188.

Taft, R.W., (1956), In: Newman, M.S., (ed.), *Steric Effects in Organic Chemistry*, Wiley, New York.

Tan, C., Chen, H., Wang, C., Zhu, W., Wu, T., and Diao, Y., (2013), *A multi-model fusion strategy for multivariate calibration using near and mid-infrared spectra of samples from brewing industry*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 105, 1-7.

Tano, K., (1996), *Multivariate Modelling and Monitoring of Mineral Processes using Partial Least Squares Regression*, Licentiate Thesis, Luleå University of Technology, Sweden.

Tano, K., (2005), *Continuous Monitoring of Mineral Processes with Special Focus on Tumbling Mills*, PhD Thesis, Luleå University of Technology, Sweden.

Teague, S.J., Davis, A.M., Leeson, P.D., and Oprea, T., (1999), *The Design of Leadlike Combinatorial Libraries*, Angewandte Chemie International Edition, 38, 3743-3748.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., and Lauro, C., (2005), *PLS Path Modelling*, Computational Statistics & Data Analysis, 48, 159-205.

Teppola, P., Mujunen, S.P., Minkkinen, P., Puijola, T., and Pursiheimo, P., (1998), *Principal Component Analysis, Contribution Plots and Feature Weights of Sequential Process Data from a Paper Machine's Wet End*, Chemometrics and Intelligent Laboratory Systems, 44, 307-317.

Teppola, P., and Minkkinen, P., (2000), *Wavelet-PLS Regression Models for Both Exploratory Data Analysis and Process Monitoring*, Journal of Chemometrics, 14, 383-400.

Teppola, P., and Minkkinen, P., (2001), *Wavelets for Scrutinizing Multivariate Exploratory Models Through Multiresolution Analysis*, Journal of Chemometrics, 15, 1-18.

Thomas, V., Robert, S., and Richard, J., (2008), *DiPCA_Cluster: An Optimal Alternative to DiPLS_Cluster for Unsupervised Classification*, Chemometrics and Intelligent Laboratory Systems, 90, 8-14.

Tian, X., Chen, G., and Chen, S., (2011), *A Data-based Approach for Multivariate Model Predictive Control Performance Monitoring*, Neurocomputing, 74, 588–597.

Tomasi, G., van den Bergh, F., and Andersson, C., (2004), *Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data*, Journal of Chemometrics, 18, 231-241.

Tompson, C.J.S., (1990), *The Lure and Romance of Alchemy - a History of the Secret Link Between Magic and Science*, Bell Publishing Company, New York.

Topliss, J.G., and Edwards, R.P., (1979), *Chance Factors in Studies of Quantitative Structure-Activity Relationships*, Journal of Medicinal Chemistry, 22, 1238-1244.

Trygg, J., and Wold, S., (1998), *PLS Regression on Wavelet Compressed NIR Spectra*, Chemometrics and Intelligent Laboratory Systems, 42, 209-220.

Trygg, J., Kettaneh-Wold, N., and Wallbäcks, L., (2001), *2-D Wavelet Analysis and Compression of On-line Industrial Process Data*, Journal of Chemometrics, 15, 299-319.

Trygg, J., and Wold, S., (2002), *Orthogonal Projections to Latent Structures (OPLS)*, Journal of Chemometrics, 16, 119-128.

Trygg, J., (2002), *O2-PLS for Qualitative and Quantitative Analysis in Multivariate Calibration*, Journal of Chemometrics, 16, 283-293.

Trygg, J., and Wold, S., (2003), *O2-PLS, a Two-Block (X-Y) Latent Variable Regression (LVR) Method With an Integral OSC Filter*, Journal of Chemometrics, 17, 53-64.

Trygg, J., (2004), *Prediction and Spectral Profile Estimation in Multivariate Calibration*, Journal of Chemometrics, 18, 166-172.

Tysklind, M., Andersson, P., Haglund, P., van Bavel, B., and Rappe, C., (1995), *Selection of Polychlorinated Biphenyls for use in Quantitative Structure-Activity Modelling*, SAR and QSAR in Environmental Research, 4, 11-19.

Töndel, K., Vik, J.O., Martens, H., Indahl, U.G., Smith, N., and Omholt, S.W., (2013), Hierarchical multivariate regression-based sensitivity analysis reveal complex parameter interaction patterns in dynamic models, Chemometrics and Intelligent Laboratory Systems, 120, 25-41.

Undey, C., and Cinar, A., (2002), *Statistical Monitoring of Multistage, Multiphase Batch Processes*, IEEE Control Systems Magazine, October 2002, 40-52.

Undey, C., Ertunc, S., Cümar, A., (2003), *Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis*, Industrial & Engineering Chemistry Research, 42, 4645-4658.

Uppgård, L., Sjöström, M., and Wold, S., (2000). *Multivariate Quantitative Structure-Activity Relationships for the Aquatic Toxicity of Alkyl Polyglucosides*, Tenside Surfactants Detergents, 37, 131-138.

Van der Voet, H., (1994), *Comparing the Predictive Accuracy of Models Using a Simple Randomization Test*, Chemometrics and Intelligent Laboratory Systems, 25, 313-323.

Van de Waterbeemd, H., (1995), *Chemometric Methods in Molecular Design,* In: Mannhold, R., Krogsgaard-Larsen, P., and Timmerman, H., (Eds.), Methods and Principles in Medicinal Chemistry, Vol 2, VCH, Weinheim, Germany.

Van Espen, P., Janssens, G., Vanhoolst, W., and Geladi, P., (1992), *Imaging and Image Processing in Analytical Chemistry*, Analusis, 20, 81-90.

Vanhatalo, E., (2009), *Multivariate Process Monitoring of an Experimental Blast Furnace*, Quality and Reliability Engineering International.

Van Nederkassel, A.M., Xu, C.J., Lancelin, P., Sarraf, M., MacKenzie, D.A., Walton, N.J., Bensaid, F., Lees, M., Marin, G.J., Desmurs, J.R., Massart, D.L., Smeyers-Verbeke, J., and Vander Heyden, Y., (2006), *Chemometric Treatment of Vanillin Fingerprint Chromatograms. Effect of Different Signal Alignments on Principal Component Analysis Plots*, Journal of Chromatography A, 1120, 291-298.

Viteri, M.C.V., Aguilar, L.P.F.R., Medina, C.R.A., and Sanchez, M.C., (2012), *A new approach to estimate variable contribution to Hotelling´s statistic*, Chemometrics and Intelligent Laboratory Systems, 118, 120-126.

Verron, T., Sabatier, R., and Joffre, R., (2004), *Some Theoretical Properties of the OPLS Method*, Journal of Chemometrics, 18, 62-68.

Wakeling, I.N., Morris, J.J., (1993), *A Test of Significance for Partial Least Squares Regression*, Journal of Chemometrics, 7, 291-304.

Ward, J. H., Jr., (1963), *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 48, 236–244.

Westerhuis, J., Kourti, T., and MacGregor, J.F., (1998), *Analysis of Multiblock and Hierarchical PCA and PLS Models*, Journal of Chemometrics, 12, 301-321.

Westerhuis, J.A., Gurden, S.P., and Smilde, A.K., (2000), *Generalized Contribution Plots in Multivariate Statistical Process Monitoring*, Chemometrics and Intelligent Laboratory Systems, 51, 95-114.

Westerhuis, J.A., de Jong, S., and Smilde, A.K., (2001), *Direct Orthogonal Signal Correction*, Chemometrics and Intelligent Laboratory Systems, 56, 13-25.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., and van Dorsten, F.A., (2008), *Assessment of PLS-DA Cross-validation*, Metabolomics, 4, 81-89.

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., and Faber, K., (2007), *A Randomization Test for PLS Component Selection*, Journal of Chemometrics, 21, 427-439.

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., and Trygg, J., (2008), *Visualization of GC/TOF-MS-Based Metabolomics Data for Indentification of Biochemically Interesting Compounds using OPLS Class Models*, Analytical Chemistry, 80, 115-122.

Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., Rännar, S., Sandberg, M., Kettaneh-Wold, N., and Wold, S., (1998a), *Multivariate Process and Quality Monitoring Applied to an Electrolysis Process – Part I. Process Supervision with Multivariate Control Charts*, Chemometrics and Intelligent Laboratory Systems, 42, 221-231.

Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., Rännar, S., Sandberg, M., Kettaneh-Wold, N., and Wold, S., (1998b), *Multivariate Process and Quality Monitoring Applied to an Electrolysis Process – Part II. Multivariate Time-series Analysis of Lagged Latent Variables*, Chemometrics and Intelligent Laboratory Systems, 42, 233-240.

Wikström, P.B., Andersson, A.C., Forsman, M., (1999), *Biomonitoring Complex Microbial Communities Using Random Amplified Polymorphic DNA and PCA*, FEMS – Microbiology, Ecology, 28, 131-139.

Wikström, P.B., (2001), *Biomonitoring of Complex Microbial Communities that Biodegrade Aromatics*, Ph.D. Thesis, Umeå University, Umeå, Sweden.

Wikström, M., and Sjöström, M., (2004), *Identifying Cause of Quality Defect in Cheese Using Qualitative Variables in a Statistical Experimental Design*, Journal of Chemometrics, 18, 139-145.

Winiwarter, S., Bonham, N.M., Ax, F., Hallberg, A., Lennernäs, H., and Karlén, A., (1998), *Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters – A Multivariate Data Analysis Approach*, Journal of Medicinal Chemistry, 41, 4939-4949.

Winiwarter, S., Ax, F., Lennernäs, H., Hallberg, A., Pettersson, Cu., and Karlén, A., (2003), *Hydrogen Bonding Descriptors in the Prediction of Human In Vivo Intestinal Permeability*, Journal of Molecular Graphics and Modelling, 21, 273-287.

Wise, B.M., Gallagher, N.B, and Martin, E.B., (2001), *Application of PARAFAC2 to Fault Detection and Diagnosis in Semiconductor Etch*, Journal of Chemometrics, 15, 285-298.

Wold, S., (1978), *Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models*, Technometrics, 20, 397-405.

Wold, H., (1982), *Soft Modelling. The Basic Design and Some Extensions*. In: Jöreskog, K.G., and Wold, H., (eds.), Systems Under Indirect Observation, Vol. I and II, North-Holland, Amsterdam, The Netherlands.

Wold, S., and Dunn, III, W.J., (1983), *Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for their Applicability*, Journal Chemical Information & Computer Science, 23, 6-13.

Wold, S., Albano, C., Dunn, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjöström, M., (1984), Multivariate Data Analysis in Chemistry, In: B.R. Kowalski (ed.) *Chemometrics: Mathematics and Statistics in Chemistry*, D. Reidel Publishing Company, Dordrecht, Holland.

Wold, S., Geladi, P., Esbensen, K., and Öhman, J., (1987a), *Multiway Principal Components and PLS-Analysis*, Journal of Chemometrics, 1, 41-56.

Wold, S., Hellberg, S., Lundstedt, T., Sjöström, M. and Wold, H., (1987b), *PLS Modeling with Latent Variables in Two or More Dimensions*, Proceedings Frankfurt PLS-meeting, September, 1987.

Wold, S., Carlson, R., and Skagerberg, B., (1989a), *Statistical Optimization as a Means to Reduce Risks in Industrial Processes*, The Environmental Professional, 11, 127-131.

Wold, S., Kettaneh-Wold, N., and Skagerberg, B. (1989b), *Nonlinear PLS Modelling*, Chemometrics and Intelligent Laboratory Systems, 7, 53-65.

Wold, S., (1992), *Non-linear Partial Least Squares Modelling. II. Spline Inner Relation*, Chemometrics and Intelligent Laboratory Systems, 14, 71-84.

Wold, S., Johansson, E., and Cocchi, M., (1993a), *PLS*, In: Kubinyi, H., (ed.), 3D-QSAR in Drug Design, Theory, Methods, and Applications, ESCOM Science, Ledien, pp. 523-550.

Wold, S., Jonsson, J., Sjöström, M., Sandberg, M., and Rännar, S., (1993b), *DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by Principal Components Analysis and Partial Least Squares Projections to Latent Structures*, Analytica Chimica Acta, 277, 239-253.

Wold, S., (1994), *Exponentially Weighted Moving Principal Components Analysis and Projections to Latent Structures*, Chemometrics and Intelligent Laboratory Systems, 23, 149-161.

Wold, S., Kettaneh, N., and Tjessem, K., (1996), *Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection*, Journal of Chemometrics, 10, 463-482.

Wold, S., Antti, H., Lindgren, F., and Öhman, J., (1998a), *Orthogonal Signal Correction of Near-Infrared Spectra*, Chemometrics and Intelligent Laboratory Systems, 44, 175-185.

Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A., (1998b), *Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments*, Chemometrics and Intelligent Laboratory Systems, 44, 331-340.

Wold, S., Sjöström, M., Eriksson, L., (1999), *PLS in Chemistry*, In: The Encyclopedia of Computational Chemistry, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R., Eds., John Wiley & Sons, Chichester, 1999, pp 2006-2020.

Wold, S. and Josefson, M., (2000), Multivariate Calibration of Analytical Data, in: Meyers, R., Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd., pp 9710-9736.

Wold, S., Trygg, J., Berglund, A., and Antti, H., (2001a), *Some Recent Developments in PLS Modeling*, Chemometrics and Intelligent Laboratory Systems, 58, 131-150.

Wold, S., Sjöström, M., and Eriksson, L., (2001b), *PLS-Regression: A Basic Tool of Chemometrics*, Journal of Chemometrics, 58, 109-130.

Wold, S., Berglund, A., and Kettaneh, N., (2002), *New and Old Trends in Chemometrics. How to Deal With the Increasing Data Volumes in RDP – With Examples From Pharmaceutical Research and Process Modelling*, Journal of Chemometrics, 16, 377-386.

Wold, S., Josefson, M., Gottfries, J., and Linusson, A., (2004), *The Utility of Multivariate Design in PLS Modeling*, Journal of Chemometrics, 18, 156-165.

Wold, S., Eriksson, L., and Kettaneh, N., (2010), *PLS in Data Mining and Data Integration*, Handbook of Partial Least Squares, Springer.

Worley, B., Halouska, S., and Powers, R., (2013), *Utilities for quantifying separation in PCA/PLS-DA score plots*, Analytical Biochemistry, 433, 102-104.

Wormbs, G., Larsson, A., Alm, J., Tunklint-Aspelin, C., Strinning, O., Danielsson, E., and Larsson, H., (2004), *The use of Design of Experiments and Sensory Analysis as Tools for the Evaluation of Production Methods for Milk*, Chemometrics and Intelligent Laboratory Systems, 73, 67-71.

Wu, J., Hammarström, L.G., Claesson, O., and Fängmark, I.E., (2003), *Modeling the Influence of Physico-Chemical Properties of Volatile Organic Compounds on Activated Carbon Adsorption Capacity*, Carbon, 41, 1309-1328.

Wu, H., White, M., and Khan, M.A., (2011), *QBD: An Integrated Process Analytical Technology (PAT) Approach for a Dynamic Pharmaceutical Co-precipitation Process Characterization and Process Design Space Development*, International Journal of Pharmaceutics, 405, 63-78.

Xu, H., Liu, Z., Cai, W., and Shao, X., (2009), *A Wavelength Selection Mmethod Based on Randomization Test for Near-infrared Spectral Analysis*, Chemometrics and Intelligent Laboratory Systems, 97, 189-193.

Xu, X., Costa, A.P., Khan, M.A., and Burgess, D.J., (2012), *Application of quality by design to formulation and processing of protein liposomes*, International Journal of Pharmaceutics, 434, 349-359.

Zeaiter, M., Roger, J.M., and Bellon-Maurel, V., (2006), *Dynamic Orthogonal Projection. A New Method to Maintain the On-line Robustness of Multivariate Calibrations. Application to NIR-based Monitoring of Wine Fermentations*, Chemometrics and Intelligent Laboratory Systems, 80, 227-235.

Zuegge, J., Fechner, U., Roche, O., Parrott, N.J., Engkvist, O., and Schneider, G., (2002), *A Fast Virtual Screening Filter of Cytochrome P450 3A4 Inhibition Liability of Compound Libraries*, QSAR, 21, 249-256, 2002.

Öberg, T., (2003), *Optimization of an Industrial Afterburner*, Journal of Chemometrics, 17, 5-8.

Österberg, T., and Norinder, U., (2000), *Theoretical Calculation and Prediction of P-Glycoprotein-Interacting Drugs Using Molsurf Parametrization and PLS Statistics*, European Journal of Pharmaceutical Sciences,10, 295-303.

Index

2

2-D wavelet analysis 263, 487

A

Absolute distance to the model 452
Aligned Vectors 446
All data are needed 19, 32
Alternating conditional expectations (ACE) 375
Alternative to block-scaling 369
Alternative to variable selection 368
Approaches to estimate non-linearity 385
Approaches to MSPC 283
Assignable causes 171, 189
Assumptions of PLS 86
Asymmetric case 194, 208
Autocorrelation function 309, 312, 317, 322
Automatic transformation
    Criteria 459
Automatic transformation criteria 459
Autoregressive 308
Auto-scaling 36, 244
Autoscaling 459

B

Bandpass of frequencies 259
Base level model 359, 360, 369, 370, 371
Basic assumption of SPC 267, 284
Batch
    Cross validation rules 463
Batch duration and synchronization 305
Batch level 291, 296, 299, 301
Batch processes 88, 173, 189, 288, 289, 473, 481, 490
Batch project
    Cross validation rules 463
Batch statistical process control 4, 171, 283, 287
Batch variable importance 302, 304
Batch vectors 446
Battery-scaling 248
Benefits of transformation 235
Best basis 260
Bimodality 383
Block scaling 459

Blocks of logically related variables 355
Blocks of variables 369
Block-scaling 369
Block-scaling 247, 253

C

CA 430
Calculation
PLS-Trees 432
Scaling weight 460
Categorical variable 18, 373, 377, 383
Causal relationship 20, 295
Causation 20, 32
Chained filters 259
Challenges of multivariate data 19
Class
Scaling 460
Classical methods of statistics 22
Classification of observations 192, 194, 209
Classification rate 209
Cluster analysis
Hierarchical - HCA 431
Overview 430
PLS-Trees 431
References 432
Cluster-based design 136, 474
CoeffC 458
CoeffCS 457
Coefficients
Background 456
CoeffC 458
CoeffMLR 457
Rotated 458
Scaled and centered 457
Unscaled 457
Y-Related Profiles 430
CoeffMLR 457
CoeffRot 458
COMFA 472
Compare PLS with OPLS/O2PLS 429
Concept of a model 17, 28, 407
Conceptual basis of semi-empirical modeling 30
Confidence intervals 456
Confidence Intervals 31, 73, 85, 408, 419, 424
Confounding and/or undesirable variation 369
Continuous variables 18, 383
Contribution plot 358, 361, 362
Contribution tool 360
Contribution-based diagnostics 176
Control chart
Nomenclature and notation 466
Statistics 466
Target and standard deviation 467
Controlled factors 18

Coomans plot 199, 206, 208
Correction factor 450, 452
Correlation and causation 20, 32
Correspondence analysis 54, 251, 252
Cosine of the angles 43
COST approach 17, 21, 32
Criteria automatic transformation 459
Critical distance to the model 454
Cross validation
    What is 461
Cross validation rules
    Batches 463
    Non Significant 463
    Significant component 462
Cross-correlation function 318, 322
Cross-validation (CV) is a practical 52
Cross-validation is known to face problems 421
Cumulative explained variation 81
Curved (non-linear) relationships 64
CuSum chart 273, 276, 282, 394
CV-ANOVA 85
    Background 464
Cylindrical tolerance volume 85

D

Data correction and compression 34
Data preparation and standardization 4
DCrit 454
Degrees of non-linearity 374, 384
Dendrogram
    Background 430
Derivation of principal properties 133
Design of experiments 13, 17, 20, 22, 239, 241, 295, 301, 376, 474
Deterministic component 17
Deviating time points 62
Diagnose the process upset 361
Diagnosing and handling non-linearity 375
Digitization 150
Dimensionality problem 23
Discontinuity 383
Discrete factor 132
Discrete wavelet transform 260
Discriminant analysis 1, 4, 88, 192, 203, 206, 208, 210, 477, 481
Discriminant Analysis models
    Cluster Analysis 430
Disjoint class modelling 209
Disjoint principal component 196
Distance to model
    Absolute 452
    Background 452
    DModX predictionset 453
    Normalized 453
Distance to the model 8, 47, 78, 180, 197, 209, 263, 293, 382
Distinction between strong and moderate outliers 53

Distribution of the residuals 235, 236
Diverse sub-set 132
DMod
Absolute 452
Normalized 453
Predictionset 453
D-optimal design 135
Double-centering 251

E

Effect of mean-centering 38
EWMA chart 275, 282, 394
Expanded terms scaling 460
Expansion of the X-matrix 233, 239, 242, 385
Experimental objectives 140
Expert systems 284
Explained variance 49, 81
Explained variation of a variable 49
Exploring process memory 275
Exploring time series properties of scores 181
Exponentially decreasing weights 275
Extensions of PLS 88
External RMSEP 380
External validation 156, 165, 384, 419, 423

F

Factor analysis 1, 54, 305
Feed-back controllers 190
Feedback process control 284
Feed-back systems 190
Filtering of data 369
Final conditions of new batches 298
Finite duration process 287
Finite impulse response (FIR) model 311, 320
First PLS component 58, 64, 65, 77, 81, 247, 375
First-order filter EWMA 275
Fisher's Exact test 466
Fit
Methods background 425
Formulas and descriptions 448
Formulating rules of classification 192
Four levels of PAT 333
Four levels of pattern recognition 193, 469
Fractional factorial design 135
Full rank 1, 20

G

Genetic algorithms (GA) 375
Geometric interpretation of PCA 37
Geometry of PLS 57, 63, 484
GIFI-PLS 373, 375, 377, 378, 379, 380, 381, 384, 385
GOLPE 470
Goodness of fit 50, 52, 82, 419
Goodness of prediction 50, 52, 82, 155, 419

GRID 472
Group contribution plot 202

H

Hard block-scaling 248
HCA 430, 431
Hierarchical
    CA 431
    Cluster Analysis 431
Hierarchical Cluster Analysis 431
Hierarchical modelling 355
Hierarchical PCA and PLS 88, 355
Hierarchical PLS and PCA 88
Higher-order components 43, 51, 66, 165, 185
Horseshoe problem 252
Hotelling´s T^2 tolerance ellipse 361
Hotelling's T2 450
Hotelling's T2Range
    Calculation 450
How many PLS components are really necessary 66

I

Identification of transfer function model 312
Identifying contribution from variables 180
Identifying the time series model 310
Implementing CuSum 273
Implementing EWMA 275
Implicit non-linear latent variable regression, INLR 376
Implicit non-linear latent variable regression, INLR 240, 376
Inclusion/exclusion of variables 244
Increased interpretability 368
Information scaling 256, 262
Inner relation 64, 69, 77, 86, 88, 141, 184, 318, 375–76, 379, 382, 385, 415, 489
Inspection of time series properties 409
Inspection of univariate statistics 409
Intelligently selected observations 1
Interpret the score plot 42
Interpretation of refined model 411
interpreting deviations 268
intrinsic molecular properties 132
Introduction to GIFI-PLS 377
Iterative nature of MVDA 408, 424

J

Jack-knifing 456

K

Karhunen-Loeve expansion 54
K-dimensional space 1, 28, 33, 37, 39

L

Lag
    Scaling 461

Vectors 445
Latent structure 240, 242
Latent variable (LV) regression 373
Latent variable models 2, 88, 149
Latent variable regression 2, 73, 240, 376, 470
Lead optimization 374
Leave-one-out approach 84
Leverage 46, 189, 253
Leverages 449
Limitations of traditional calibration 148
Linear combinations 30, 53, 70, 85, 87, 312, 414
Linear discriminant analysis 1, 213
Lipophilicity 86, 132, 134, 138, 142, 473
Log-transformation 77, 233, 375, 410
Low-dimensional plane 33

M

Main steps of GIFI-PLS 377, 384
Main steps of multivariate calibration 147, 151
Main steps of multivariate characterization 132, 144
Main steps of MVDA 407
Matrix of dummy variables 203, 209
Maximum covariance 69, 414
Maximum variance direction 40
Maximum variance least squares projection 69
Mean-centering 34, 37, 38, 52, 57, 58, 63, 84, 153, 209, 233, 244, 246, 249, 253
Megavariate analysis 1, 96, 118, 330, 356
Membership significance level 455
Memory of the process 275
Method of quantifying discrete changes 131
Mild non-linearity 374, 385
Missing data 1, 7, 20, 22, 32, 56, 140, 171, 252, 257, 265, 288, 304, 370, 410
Missing value
Correction factor 450
Missing values correction factor 450
Model concept 28, 32
Model diagnostics 50, 76, 82
Model plane 27, 41, 43, 47, 164, 180
Model systems 132, 137
Model transparency 368
Modeling power 449, 454
Moderate outliers 46, 53, 78, 87, 173, 177, 179, 184, 189, 415
Modifications, GIFI I – GIVI IV, of the methodology 386
Monitoring 361
Monitoring the state of the process 172
Mother wavelet 259
MPow 449, 454
MPow weighted DMod 449, 454
MPowX 449, 454
MTSA 4, 307, 312, 314, 315, 317, 321
Multicollinearity 20, 23, 32, 88, 135, 213
Multidimensional space 23, 25, 32, 41
Multiple linear regression 1, 20, 22, 87
Multiple responses 63

Multiplicative signal correction (MSC) 255
Multiresolution analysis 260
Multivariate adaptive regression splines (MARS) 375
Multivariate calibration 2, 13, 37, 56, 57, 77, 88, 147, 152, 161, 163, 166, 171, 233, 246, 255, 256, 322, 373, 469, 474, 481, 485, 490
Multivariate characterization 1, 2, 131, 137, 143, 171, 203, 248, 476, 484
Multivariate classification 1, 4, 208
Multivariate CuSum charts 274
Multivariate data analysis cycle 408
Multivariate design 135, 139, 144, 423, 474, 479
Multivariate EWMA charts 275
Multivariate normality 46
Multivariate process modelling 4, 171, 174, 189, 265
Multivariate Shewhart charts 270
Multivariate statistical process control 4, 171, 265, 269, 287, 483
Multivariate time series analysis 4, 307, 322
Multi-way techniques 283

N

Necessary condition for PLS-DA 209
Negative logarithm 29, 234
Negatively (inversely) correlated 42
Neural networks (NN) 375
NIPALS 56, 257
No scaling 37, 243, 253
Nomenclature
    Control charts 466
Non Significant component 463
Non-linear PLS 88, 374, 375, 385, 471, 484
Non-linear PLS 373, 375, 385
Non-linearity 382
Non-linearity between X and Y 77, 414
Normal distribution 31, 234
Normal probability plot of residuals 236
Normalized DMod 453
Notation control charts 466
Notation used in PCA 33
Number of CV groups 84

O

O2PLS
    Background 427
    Hierarchical approach 429
    Orthogonal component 429
    Predictive component 429
    References 430
    with hierarchical 429
Objectives of classification 208, 230
Observable group similarity 209, 231
Observation diagnostics 46, 76
Observation Risk 468
Observations
    Risk 468
Observations and Loadings vectors 438

Observations are paired 249
Occurrence of a special event 267, 284
One or several response variables 86
One-variable-at-a-time 21
On-line PAT 334
On-line predictions 15, 187, 304
OOC plot 446
OPLS 89, 90, 96, 97, 98, 100, 101, 107, 108, 112, 113, 114, 115, 117, 118, 119, 120, 122, 123, 125, 126, 128, 130, 427
OPLS/O2PLS
    Background 427
    Orthogonal component 429
    Predictive component 429
    References 430
    with hierarchical 429
OPLS-DA 89, 215, 216
Optimal model dimensionality 52
Orientation of the model plane 43
Orientation of the obtained plane 27
ORisk 468
Orthogonal components 108, 112, 113, 114, 130
Orthogonal PLS 89
Orthogonal PLS 89, 117
Orthogonal PLS modeling
    Background 427
    Orthogonal component 429
    Predictive component 429
    References 430
    with hierarchical 429
Orthogonal signal correction, OSC 257
Out Of Control Summary 446
Outlier handling 370

P

PARAFAC (parallel factor analysis 305
Parametric release 323, 334
Pareto scaling 243, 246, 253
Partial least squares projections to latent structures 5, 55, 255, 489
Partial Least Squares Projections to Latent Structures 426
Path models 56
Pattern recognition 6, 192, 469, 486
PCA 425
PCA diagnostics 46
PCA for data overview 411
PCR 87
Peptide QSAR 249, 378, 380
Philosophy of local modelling 368
Philosophy of model building 17
PLS
    Time series analysis 464
    Trees 431
    Vs. OPLS/O2PLS 429
    What is? 426
PLS prediction procedure 85

PLS regression coefficients 72, 207, 310, 313
PLS score plots 67, 76, 141, 328, 385
PLS time series analysis 307, 310, 319, 321
PLS vs OPLS and O2PLS 429
PLS weights 56, 66, 70, 74, 141, 186, 245, 257, 292, 302, 318, 321, 328, 375, 379–80, 415
PLS with multiple responses 63
PLS-DA 88, 203
PLS-Trees
    Calculation 432
    What is? 431
Positively correlated 20, 42, 292
Practical rank 2, 87
Prediction intervals in PLS 86
Predictions for new observations 137
Predictions of new observations 419, 424
Predictionset
    DMod 453
Predictive component 96
Predictive EWMA 275
Predictive residual sum of squares 52, 84, 420
Predictive validation 409, 419, 423
Preference mapping 482
Pre-processing of data 57, 209, 252
Pre-requisites of PAT 332
Preserving batch direction 296
Preserving variable direction 291
Pre-treatment of data 34
Principal Component Analysis 1, 4, 5, 26, 33, 53, 87, 89, 117, 131, 137, 323, 1, 425, 469, 478, 487
Principal component loadings 42
Principal Component modeling 425
Principal component regression 87
Principal component regression, PCR 371
Principal component scores 45, 307, 322
Principal components analysis 6, 26, 477, 489
Principle of analogy 191
Principles of projections 17, 23, 28, 32
Problems of process data 171, 265, 282
Process analytical technology 4
Process Analytical Technology 323, 331
Process modeling 55, 66
Process modelling 3, 7, 171, 174, 189, 194, 204, 208, 241, 244, 265, 304, 322, 373, 412, 419
Process monitoring and optimization 56
Process state visualisation 355
Process status visualization 20, 172, 326
Process upset 356, 361, 362
Product attributes 265, 267
Projection co-ordinate 60
Projection Methods 1, 4, 5, 17, 20, 23, 30, 32, 57, 88, 89, 117, 150, 323, 1
Projection to latent structures 56
Projections in higher-dimensional space 26
Projections to Latent Structures 1, 4, 5, 55, 89, 117, 255, 323, 355
Properties of GIFI-PLS 378
Pure profile 104
Pure spectral profiles 104, 106, 111

Q

Q2 ..... 448, 449
Q2V 448, 449
Qualitative variable 18
Quality control 2, 9, 174, 419
Quantitative relation between X and Y 70, 414
Quantitative sequence-activity modelling 478
Quantitative structure-activity relationships 56, 470, 483
Quantitative variable 18

R

R2V 448
R2Vadj 448
R2X 448
R2Xadj 448
R2Y 448
R2Yadj 448
Rationale of multivariate characterization 132
Red in Transform page 459
References
    Cluster analysis 432
    OPLS/O2PLS 430
    PLS time series 464
Refined modelling with PCA or PLS 411
Representative experiments 131, 239, 376
Representative set of observations 30, 133, 135
Residual observation variance 48, 80
Residual variable variance 50, 81
Residual variable variation 50, 81
Residual vector 59, 62, 312
Response contour plot 186, 241, 328
Response contour plotting 186
Response permutation and cross-validation 421
Response permutation testing 419, 422
Ridge regression 87
Risk of spurious results 22
RMSEcv 458
RMSEE 458
RMSEP 380, 458
Root cause 358
Root mean square error of prediction 380
Rotated coefficients 458
Rotated regression coefficients 104, 106, 111, 112
Rotated regression coefficients 97, 120
RR 87
RSD 449

S

Savitsky-Golay smoothing 255
Scalability 368
Scale page
    Background 459
    Lags 461

Scaling
After changing observation selection ... 461
Classes ... 460
Expanded terms ... 460
Lags ... 461
Transformed variables ... 460
Weight calculation ... 460
Scaling and centering ... 410
Scaling of data ... 34, 233, 243, 254
Scaling to unit variance ... 36, 39, 53, 57, 63, 209, 233, 243, 245, 247, 253
Scope of SPC ... 268
Score
Vectors ... 451
Score vector ua ... 69
Score vectors ta ... 69
Second PLS component ... 65, 71, 83, 247
Secondary variation in X ... 90, 113
Semi-empirical modeling ... 29
Sensitive to scaling ... 37
Sequence data ... 478
Set point ... 244, 374
Shewhart chart ... 270, 272, 276, 277, 413
Signal compression ... 233, 255, 257, 260
Signal correction ... 57, 157, 233, 255, 407, 410, 474, 488, 490
Signal correction and compression ... 255, 256, 262, 410
Significance level membership ... 455
Significant component ... 462
SIMCA ... 2, 47, 48, 52, 57, 69, 70, 97, 113, 118, 120, 121, 165, 196, 238, 258, 291, 300, 354, 389, 399, 411, 433
SIMCA method ... 197
SIMCA methodology ... 197
SIMCA offline ... 279
SIMCA-control ... 352
SIMCA-online ... 279
SIMCA-online software ... 279
Similarity and class ... 191
Simplified the response function ... 236
Skewness of each variable ... 237
S-line ... 215, 218, 220, 222
Soft block-scaling ... 248
Soft independent modeling of class analogy ... 196
Space filling designs ... 144
Spectral data ... 14, 61, 71, 153, 233, 255, 260
Spline PLS (SPLS) ... 375
S-plot ... 215, 220, 221, 468
Standard curve ... 32, 147
Standard error ... 456
Standard normal variate (SNV) correction ... 255
State estimation ... 284
Stationary ... 132, 144, 307, 309, 313, 318, 322
Statistical notes ... 425
Steady-state or dynamic conditions ... 173
Step Response Plot ... 320
Stochastic component ... 17

Strong non-linearity 374–75
Strong outliers 46, 53, 161, 173, 177, 189, 253, 271, 320, 412
Structured noise 100, 104, 107
Structured noise 90
Subgrouping 272, 282
Super variables 355
Supervision of existing processes 172
Support vector machines (SVM) 375
SUS-plot 215, 222, 403

T

Target
    Control chart 467
Target directed filtering 355
Taylor series expansions 29
Temporary process upsets 48, 189
The asymmetric case 194, 208
Theoretical foundation 31, 57, 407
Theoretical models 29
Three-factor mixture design 163
Time Series
    Analysis 464
    References 464
Time series data 260, 307, 310
Tolerance volume 79, 85, 194, 205
Tolerance volumes 78, 149, 194
Top level PLS model 359
Training set selection 137, 139
Transfer function models 311
Transform page coloring criteria 459
Transformation of variables 233, 236, 242
Transformations 34, 57, 204, 234, 238, 242, 262, 409
Transition from SPC to MSPC 269, 283
Trimming 234, 243, 252, 409
Two levels of batch modelling 291
Types of data 6, 16, 17, 190
Typical classification situations 191

U

Uncertainties 456
Uncontrolled factors 18
Uncorrelated responses 73, 74, 416
Undesirable feature 369
Unexplained variation 78
Unified description of classical classification methods 213, 483
Unit variance (UV) scaling 36
Univariate index 174
Unrealistic model 49

V

Validate option in SIMCA 421
Variable
    Importance 455
Variable diagnostics 49, 50, 76, 80

Variable influence on projection, VIP 74
Variance-covariance matrix 1, 88
Vector
    Aligned 446
    Batch 446
    Function of component 443
    Function of lags 445
    Observations and Loadings 438
    Variables and scores 433
Vectors available 433
VIP
    Background 455
Visualization of the correlation structure 64

W

Wavelet analysis 233, 255, 256, 474, 487
Wavelet packet transform 260
Weighted averages 53, 87, 269
Window into the three-dimensional space 26
Winsorizing 234, 243, 252, 409

X

X and Y co-ordinate systems 58
X-weight vectors wa 69

Y

Y-orthogonal variation 90, 115
Y-Related Profiles 430
Y-weights ca 69

Z

Zeroing 256
Zoom-in/zoom-out capability 368
Zoom-in/zoom-out functionality 361
Zooming-in on highlighted samples 175
Z-scales 248