Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning Lesson #16 – Flipped lecture

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Flipped learning procedure

- This is a **flipped classroom**
  - the learning procedure is inverted
- Please, complete the following procedure:
  1. **attend this video lesson**
  2. **self-assess your learning**
  3. **re-read the following papers** available in the “Suggested reading” in Moodle:
     - Geladi, P., Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta,* **185**, 1–17
     - Wise, B.M., Gallagher, N.B. (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control,* **6**, 329–348
  4. **prepare questions** and anything you need to discuss with the teacher and your mates in the next lecture
- The following lesson will be:
  - 1/2 Q&A: questions (of the students) and answers (of the teacher)
  - 1/2 we will begin the last part of the course on DoE

Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #16 – Part 1

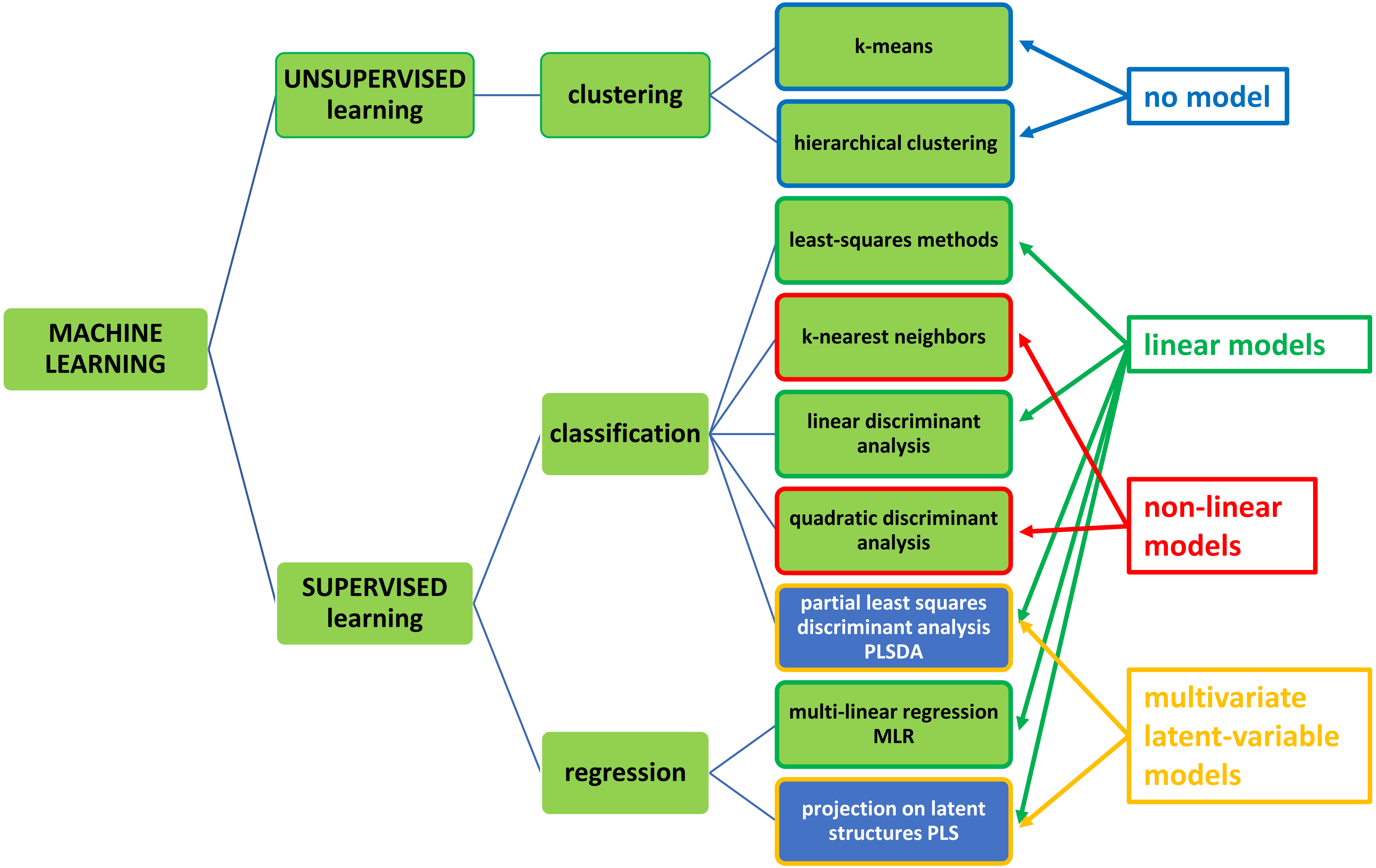
Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Recap and contextualization



What is wrong with MLR? (1/2)

- Consider the problem of **multiple linear regression MLR** with:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - the solution is: $\beta = ?$

What is wrong with MLR? (1/2)

- Consider the problem of **multiple linear regression MLR** with:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - the solution is: $\beta = 1$

What is wrong with MLR? (1/2)

- Consider the problem of **multiple linear regression MLR** with:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - the solution is: $\beta = 1$

- If we pass to a multidimensional space problem:

$$\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - a good solution can be: $\boldsymbol{\beta} = [? \quad ?]$

What is wrong with MLR? (1/2)

- Consider the problem of **multiple linear regression MLR** with:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - the solution is: $\beta = 1$

- If we pass to a multidimensional space problem:

$$\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - a good solution can be: $\boldsymbol{\beta} = [0.5 \quad 0.5]$

What is wrong with MLR? (2/2)

- If we change only a little bit the problem:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3.000001 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - a good solution can be: $\boldsymbol{\beta} = [1 \quad 0]$

- If we change the problem in another way:

$$\mathbf{X} = \begin{bmatrix} 0.999999 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

  - a good solution can be: $\boldsymbol{\beta} = [0 \quad 1]$

The solution is the same, but the regression parameters are completely different!!!
MLR is not robust because small changes in the variables generate big changes in the solution (due to collinearity)

Issues in multiple linear regression

- Consider the problem of **multiple linear regression MLR** as a premise. To relate N samples of V variables collected in a matrix $[N \times V]$ of primary variables $\mathbf{X}$ to a matrix $[N \times M]$ of M secondary variable $\mathbf{Y}$, a linear polynomial function may be assumed:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}$$

  - $\mathbf{F}$ is the $[N \times M]$ residual matrix
  - $\mathbf{B}$ is the $[V \times M]$ matrix of regression coefficients

- Three different situations may arise:
  - if $V > N$ (the number of variables is higher than the one of the samples), an infinite number of solutions exists
  - if $V = N$ (the number of variable is equal to the one of the samples), the problem has a single solution
  - if $V < N$ (the number of samples is higher than the one of the variables), an analytical solution cannot be found, but the least-squares method can be used instead:

$$\mathbf{B} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

  this $\mathbf{B}$ matrix **can be ill-conditioned if data are correlated and the solution cannot be calculated**
    - the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ in the right-hand side of the equation is not invertible

- **PLS** can be adopted to solve this regression problem:
  - to overcome the issue of **variable collinearity**
  - to investigate the **correlation between primary and secondary variables**

Projection on Latent Structures, PLS

Partial Least Squares

Partial Least Squares PLS

- Projection on Latent Structures (a.k.a. Partial Least Squares) **PLS** is a linear multivariate statistical method that, in its simplest form, relates two data matrices:
  - $\mathbf{X}$ [N observations $\times$ V variables]
    - usually easy-to-measure variables
    - online variables
    - collected frequently
  - $\mathbf{Y}$ [N corresponding observations $\times$ M variables]
    - quality variables
    - response variables
    - difficult/expensive to measure
    - collected at a low frequency
- PLS is a **linear regression technique** for the association between $\mathbf{X}$ and $\mathbf{Y}$
  - exploits the typical ability of the multivariate projection methods to analyze many **noisy and collinear data**, dealing with the *ill-conditioned regression problems*

- Geometrically, PLS finds lines/planes/hyperplanes of the closest fit for a system of points in the space of $\mathbf{X}$ that are most related and predictive for the space of $\mathbf{Y}$

PLS common applications

- PLS is commonly applied to different fields:
  - **estimation/prediction** as in soft sensing
  - **process monitoring**
  - **design and transfer of processes and products** between different scales and production sites
  - **process and product optimization**
  - **response surface modelling** in DoE
  - **QSAR**, Quantitative Structure-Activity Relationship modelling
  - **instrumentation calibration** for example in Near Infrared Spectroscopy

Mathematical formulation of PLS (1/4)

- PLS is a method which explains the **directions of maximum variability of $\mathbf{X}$ that best predict $\mathbf{Y}$**
  - PLS reduces the dimension of the system, simultaneously finding the space of LVs that are more predictive for the secondary variables and are near to the direction of maximum variability of the primary variables
- The method consists of two outer relations and one inner relation:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a^{\mathrm{T}} + \mathbf{F}$$

$$\mathbf{u}_a = b_a\, \mathbf{t}_a$$

non-linear versions of PLS can be built changing this equation

where:
  - $\mathbf{T}$ and $\mathbf{U}$ are the scores
  - $\mathbf{P}$ and $\mathbf{Q}$ are the loadings
  - $\mathbf{E}$ and $\mathbf{F}$ are the residuals that are minimized in the least squares sense
  - b_a are the **regression coefficients**: $b_a = \frac{\mathbf{u}_a^{\mathrm{T}}\mathbf{t}_a}{\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a}$
    - define a **linear relation** among the score space of the $\mathbf{X}$-scores and the scores of the $\mathbf{Y}$-space

- A smart and fast manner to calculate recursively the PLS parameters is the NIPALS algorithm

Mathematical formulation of PLS (2/4)

- PLS finds a *transformation of the* $\mathbf{X}$ *data in order to maximize the covariance of its latent variables (LVs) with the* $\mathbf{Y}$ *dataset*
- **For the first LV** this is represented by the following optimization problem:

$$\max_{\mathbf{w}_1^*}\left(\mathbf{w}_1^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1^*\right)$$
$$s.t. \quad \mathbf{w}_1^{*\mathrm{T}}\mathbf{w}_1^* = 1$$

  - it maximizes the covariance of the data projections
  - $\mathbf{w}_1$ is the $[M \times 1]$ **weight vector** and represents the coefficient of the linear combination of $\mathbf{X}$ determining the scores:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1^*$$

Mathematical formulation of PLS (3/4)

- The decomposition problem could be **solved iteratively** deflating the value of the original matrices $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X}_{a+1} = \mathbf{E}_{\mathrm{a}} = \left(\mathbf{I}_N - \frac{\mathbf{t}_a \mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}\right) \mathbf{X}_a$$

$$\mathbf{Y}_{a+1} = \mathbf{F}_{\mathrm{a}} = \left(\mathbf{I}_N - \frac{\mathbf{t}_a \mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}\right) \mathbf{Y}_a$$

- In this iterative procedure the **loadings vectors** are calculated as follow:

$$\mathbf{p}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}} \mathbf{X}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}$$

$$\mathbf{q}_a^{\mathrm{T}} = \frac{b_a \mathbf{t}_a^{\mathrm{T}} \mathbf{Y}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}$$

- To complete the formulation:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^*$$

$$\mathbf{W}^* = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}$$

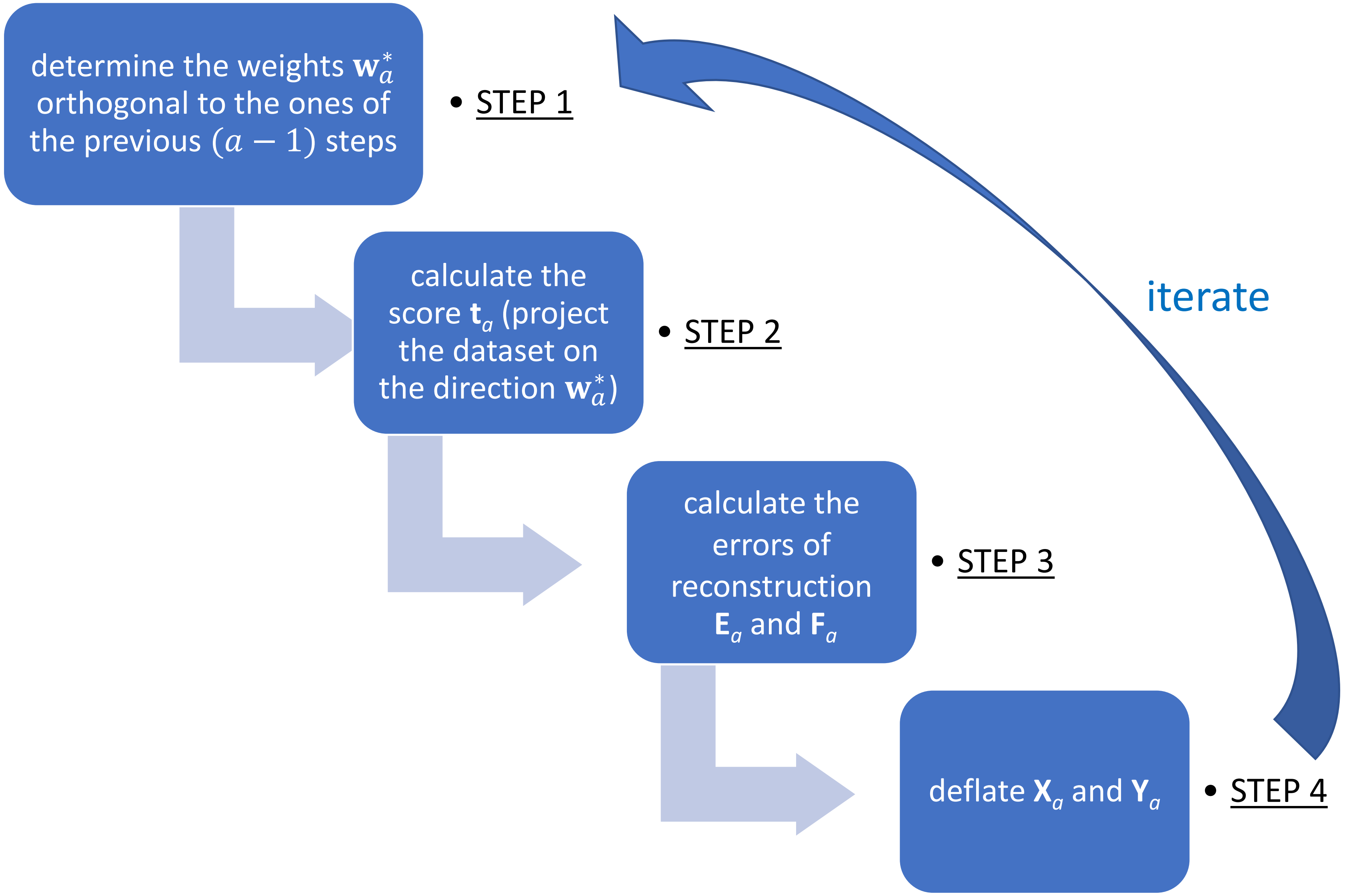
Mathematical formulation of PLS (4/4)

- The analytical solution of the abovementioned maximization problem is the same as the solution of the following eigenvector problem:

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

  - the weights $\mathbf{w}_1$ corresponds to the **eigenvector** of the covariance/correlation matrix of $\mathbf{X}$ and $\mathbf{Y}$
    - correlation if they are autoscaled
  - λ_1 is the **eigenvalue** associated to the eigenvector $\mathbf{w}_1$
    - suggestion for todays homework:
      - verify if the solution of the NIPALS algorithm is equivalent to the one of the eigenvector formulation

Latent variables extraction procedure



Summary on the mathematical formulation of PLS

- PLS describes the original datasets $\mathbf{X}$ and $\mathbf{Y}$ with a low-dimensional latent space of A LVs
  - since the variables of $\mathbf{X}$ and $\mathbf{Y}$ are correlated, these matrices are not really full rank
  - the joint variability of $\mathbf{X}$ and $\mathbf{Y}$ can be represented with a number of LVs $A \ll \min(N, V, M)$:
    - two or more correlated variables identify a common direction of variability
    - a single PC will therefore capture the variability of a lot of variables
    - a single PC is represented by the outer product of scores and loadings

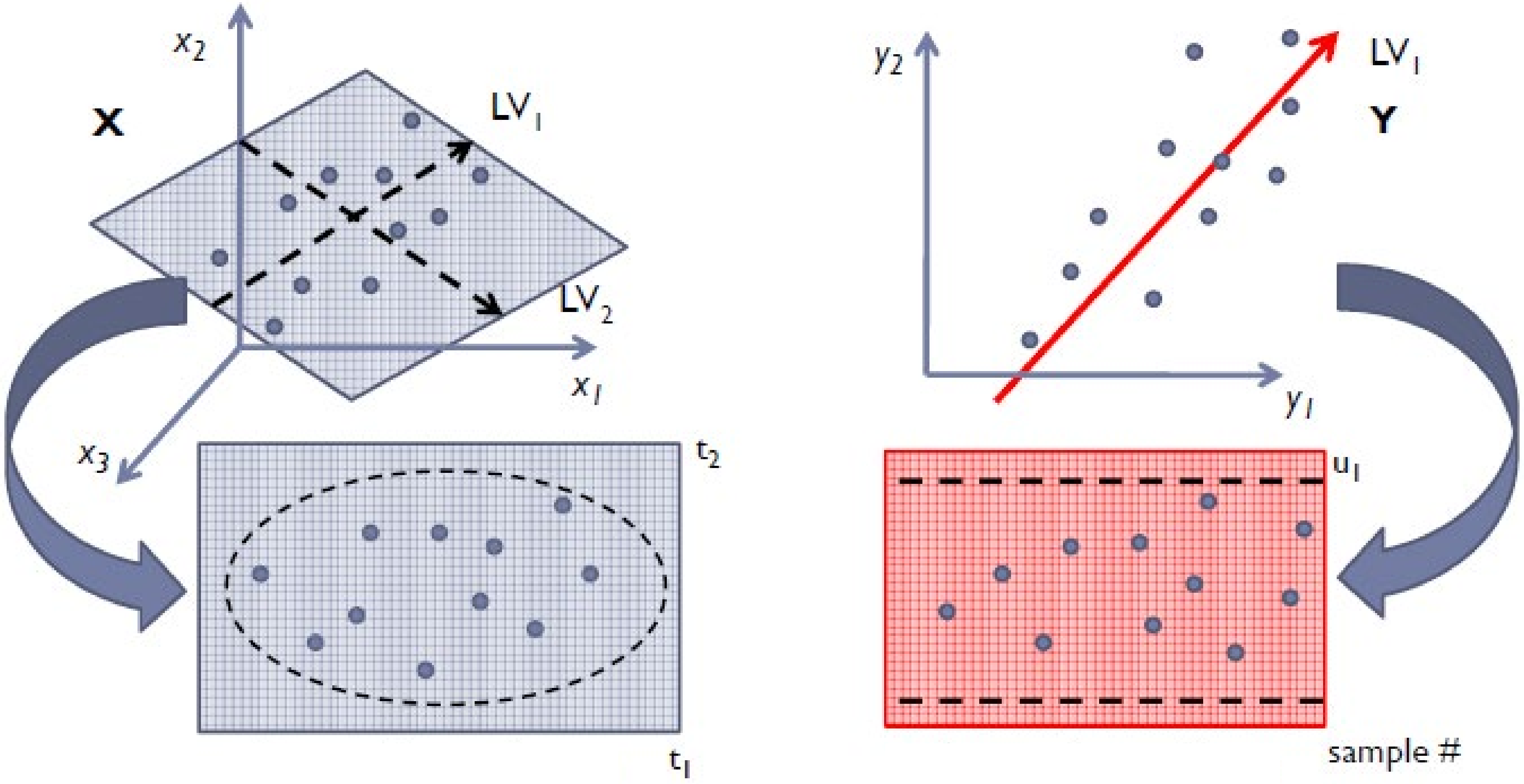
$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a^{\mathrm{T}} + \mathbf{F}$$

$$\mathbf{u}_a = b_a \, \mathbf{t}_a$$

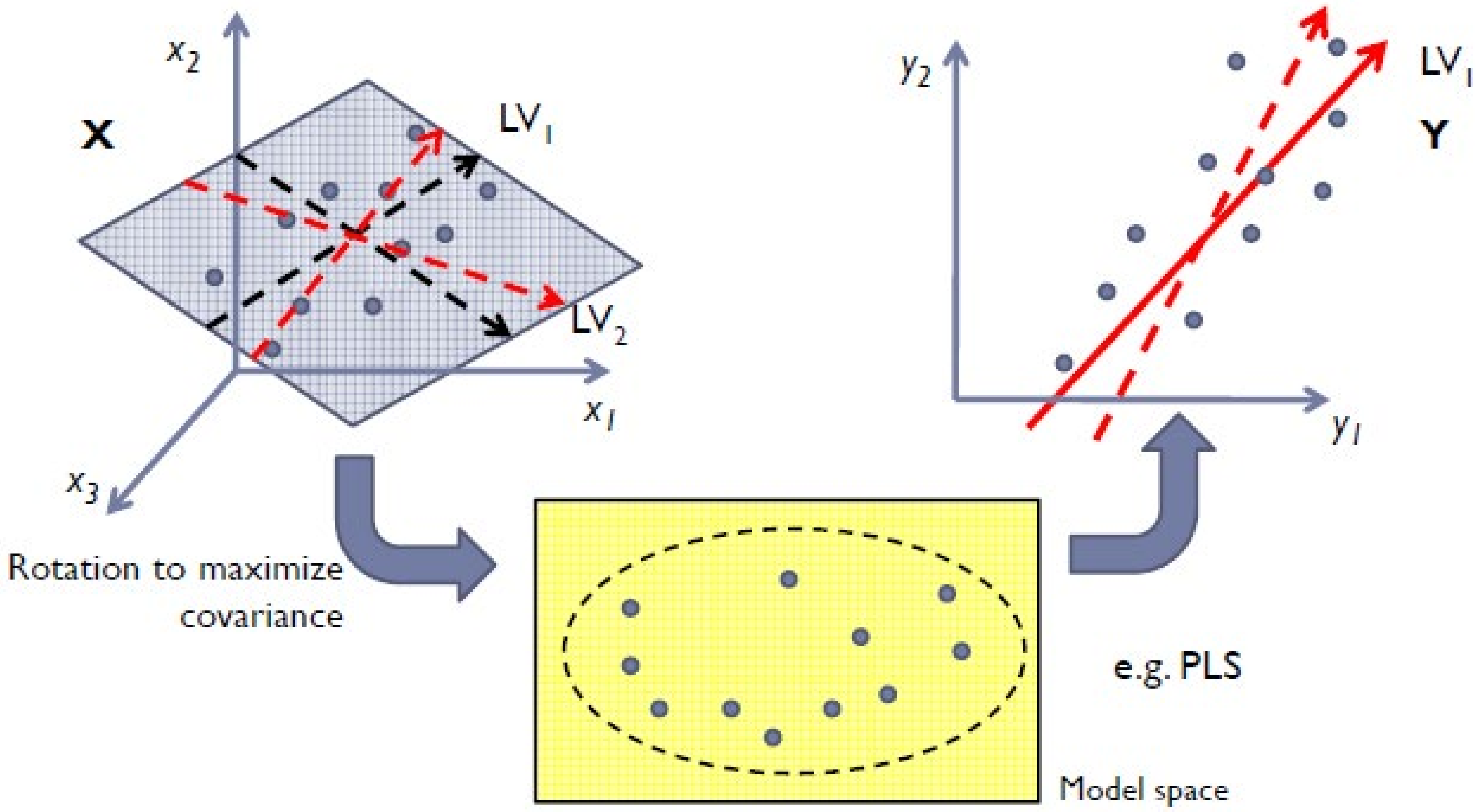
Geometrical interpretation of PLS (1/2)

- Not only PLS finds the direction of maximum variability into the $\mathbf{X}$ data…



Geometrical interpretation of PLS (2/2)

- … but also rotates them to *optimally predict* $\mathbf{Y}$



Interpretation of PLS

- $\mathbf{T}$ and $\mathbf{U}$ scores
  - projections of the observations in the space of the latent variables (i.e., the coordinates in the LV space)
  - **identify the relation among observations**
- $\mathbf{P}$ and $\mathbf{Q}$ loadings:
  - are the LVs director cosines
  - **identify the correlation between variables**

} subject to the correlation structure among $\mathbf{X}$ and $\mathbf{Y}$

- $\mathbf{E}$ and $\mathbf{F}$ residuals:
  - represent the fitting error
  - minimized in the least-square sense
  - define the distance out of the model hyperspace (i.e., the correlation structure outside the LV space)

Stop video and think to the response…

- Is PLS a regression model built on the scores of a PCA model?

Warning: PLS is not PCR!

- **PLS is not Principal Component Regression** (PCR)
  - **PCR** is a multi-linear regression model built using the scores $\mathbf{T}_{\mathrm{PCA}}$ of a PCA model built on dataset $\mathbf{X}$ as predictors of the $\mathbf{Y}$ matrix
    - the scores $\mathbf{T}_{\mathrm{PCA}}$ of a PCA model built on $\mathbf{X}$ are orthogonal
    - the regression coefficients of a MLR model can be easily estimated from the algebraic equation:

$$\mathbf{B} = \left(\mathbf{T}_{\mathrm{PCA}}^{\mathrm{T}}\mathbf{T}_{\mathrm{PCA}}\right)^{-1}\mathbf{T}_{\mathrm{PCA}}^{\mathrm{T}}\mathbf{Y}$$

      - the matrix $\left(\mathbf{T}_{\mathrm{PCA}}^{\mathrm{T}}\mathbf{T}_{\mathrm{PCA}}\right)^{-1}$ is invertible

- **PLS usually performs much better than PCR**
  - the information captured in PLS is related with the correlation $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$ of the predictor variables $\mathbf{X}$ with the outputs $\mathbf{Y}$
  - the content of information retained in PCR through $\left(\mathbf{T}_{\mathrm{PCA}}^{\mathrm{T}}\mathbf{T}_{\mathrm{PCA}}\right)^{-1}$ is “only” related with the correlation among variables within the $\mathbf{X}$ dataset
    - in PCA the covariance/correlation $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ of $\mathbf{X}$ is considered

Stop video and think to the responses…

- The meaningfulness of the regression model must be verified by answering the following questions to understand if the regression model is appropriate and reliable, as for MLR
- For PLS model:
    1. how can the **validity of the model** verified?
    2. are **all the terms of the regression model meaningful**?
    3. are the **data appropriate** to build the model?
    4. is the **linear model structure appropriate**?

PLS model verification

- The meaningfulness of the regression model must be verified by answering the following questions to understand if the regression model is appropriate and reliable:
  1. how can the **validity of the model** verified?
     - **determination coefficient R^2**
  2. are **all the terms of the regression model meaningful**?
     - it is a projection on a **reduced space** of latent variables…
  3. are the **data appropriate** to build the model?
     - the **T^2 and SPE indices** can be used for this purpose
       - defined in the same manner as in PCA, even if with different meaning
  4. is the **linear model structure appropriate**?
     - we can verify it through the **$\mathbf{t}_a$ vs. $\mathbf{u}_a$ plots** for all $a = 1, 2, \ldots, A$

Common concepts with PCA

- All the concepts utilized for PCA are also valid for PLS in terms of:
  - data **pretreatment**
    - oriented to the $\mathbf{Y}$ prediction
  - **orthogonal weights**
  - **selection of the LVs number** depends on the use of the PLS model:
    - increased $\mathbf{Y}$ variability explanation to enhance the prediction/estimation performance
    - increased $\mathbf{X}$ variability explanation to enhance the *model inversion* performance
  - **model diagnostics**
  - data **unfolding** methodologies

Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #16 – Part 2

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

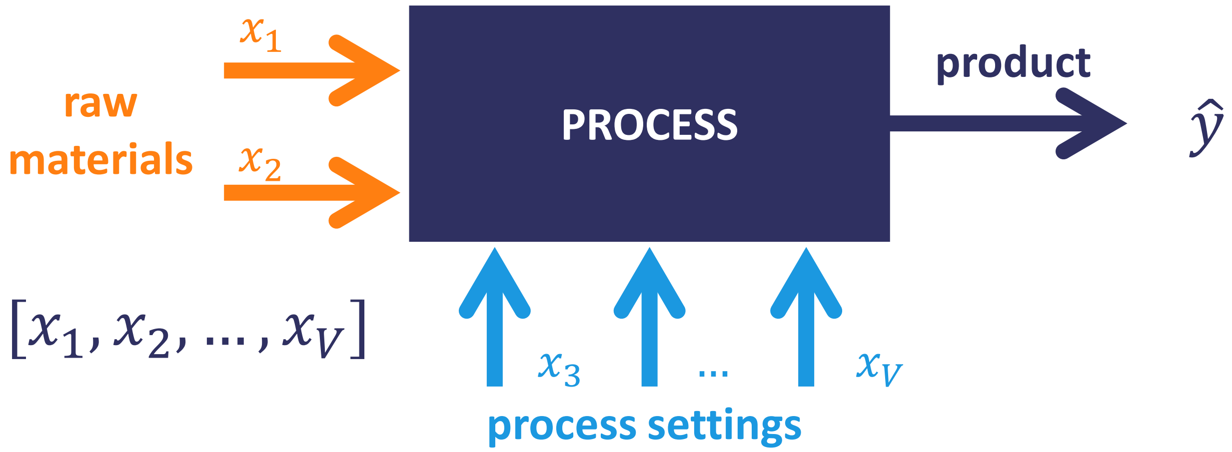
URL: https://research.dii.unipd.it/capelab/

Use of PLS

exploratory	predictive	prescriptive
understanding correlation within datasets	response estimation	design of experiments
understanding relations between datasets	class attribution in PLS Discriminant Analysis	PLS model inversion

Product quality estimation and prediction

- PLS is typically utilized in its "direct" form:
  - predicting the response, e.g. product quality
  - using the easy-to-measure inputs, e.g. raw material properties and process settings



$$\mathbf{x}_{\mathrm{NEW}} = [x_1, x_2, \ldots, x_V]$$

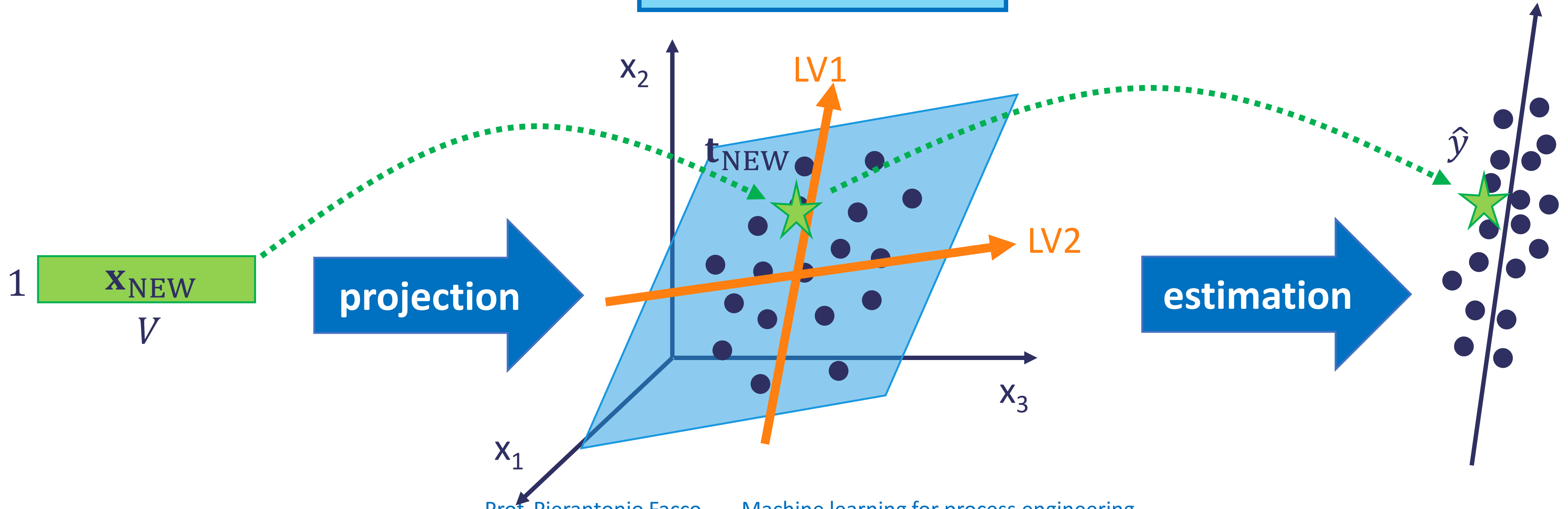
Estimations and predictions

- When a new set of predictors $\mathbf{x}_{\text{NEW}}$ is available it is possible to project it into the space of the LVs:

$$\mathbf{t}_{\text{NEW}} = \mathbf{x}_{\text{NEW}}\mathbf{P}$$

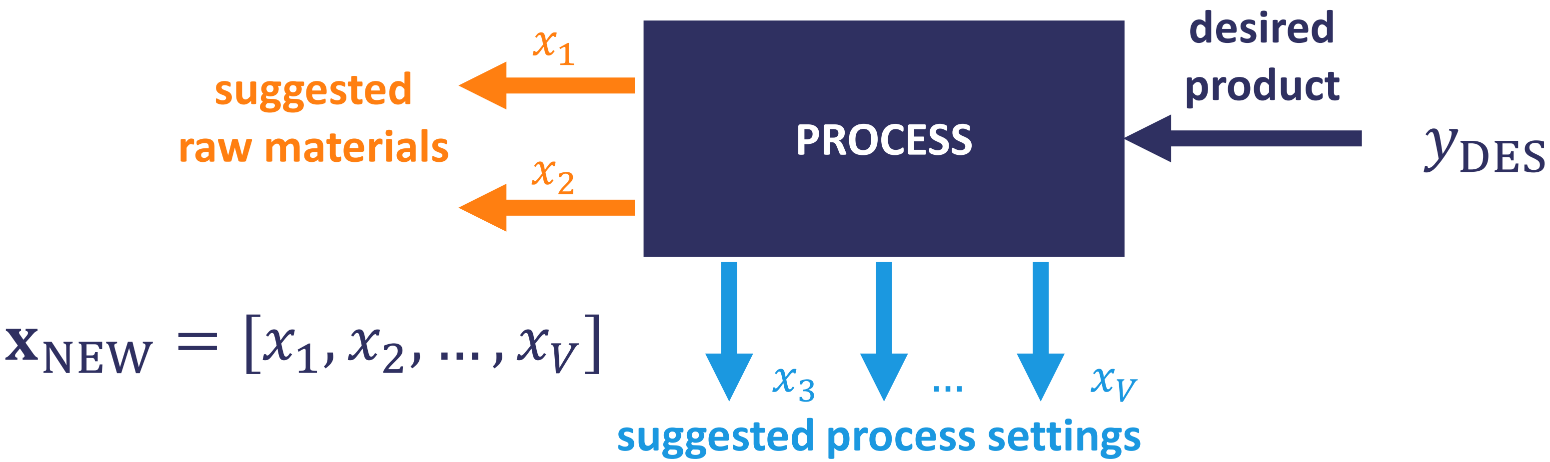
- Then it is possible to predict/estimate the response through:

$$\hat{y} = \mathbf{b}\mathbf{t}_{\text{NEW}}\mathbf{Q}^{\text{T}}$$



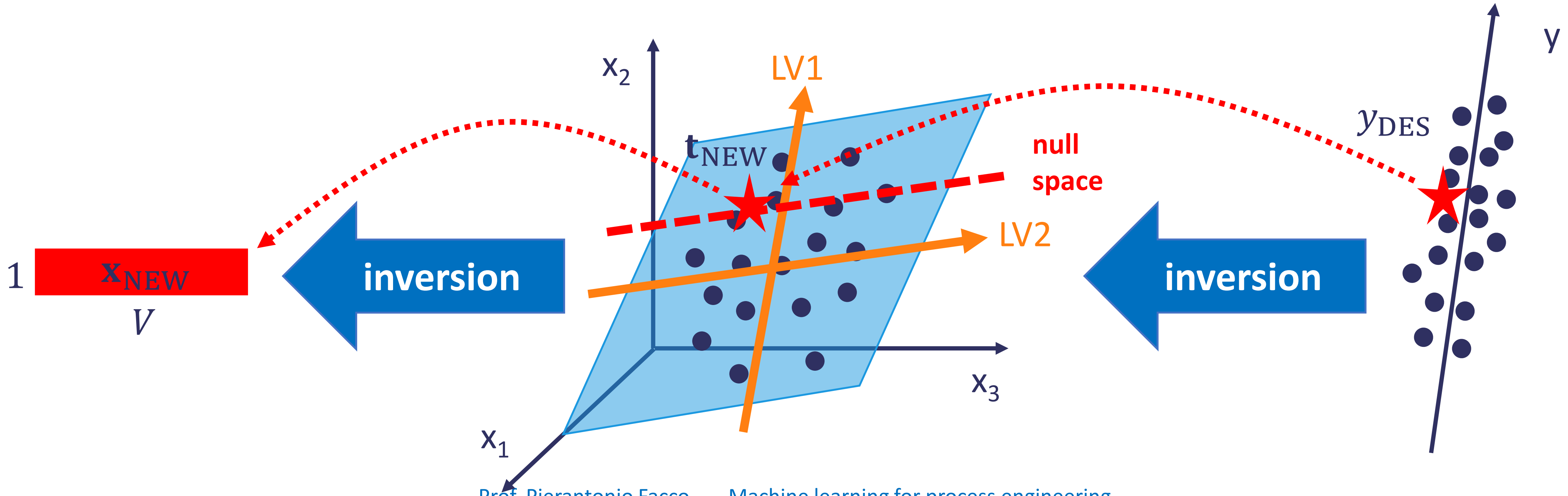
Product formulation and process design

- PLS may be inverted:
  - guaranteeing desired response, i.e., desired target quality profile for the product
  - suggesting the combination of $\mathbf{X}$, e.g. raw material properties and process settings to obtain the desired response



LV model inversion for prescriptive purposes

- PLS model inversion:
  - direct inversion: $\mathbf{x}_{\mathrm{NEW}} = \mathbf{t}_{\mathrm{NEW}}\mathbf{P}^{\mathrm{T}}$ where $\mathbf{t}_{\mathrm{NEW}} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathrm{T}}y_{\mathrm{DES}}$
  - if the real dimension of $\mathbf{X}$ is larger than the real dimension of $\mathbf{y}$ a null space exists
  - the **null space** is the locus of all the input variables $\mathbf{x}_{\mathrm{NEW}}$ corresponding to the same desired response y_{DES}



PLS in Matlab®

- PCA can be managed with several commands in Matlab®:
  - `pls` (PLS_Toolbox, Eigenvector Research Inc.)
  - `plsregress`
  - very easy NIPALS code…

- Also in this case, we will use PLS_Toolbox® graphic user interface for the sake of simplicity

PLS applications

Partial Least Squares

Example #1: exploratory analysis

exploratory
- understanding correlation within datasets
- understanding relations between datasets

predictive
- response estimation
- class attribution in PLS Discriminant Analysis

prescriptive
- design of experiments
- PLS model inversion

Example of mobile phones plastic covering

Plastic covering of mobile phones

- AIM: optimizing the polymer characteristics:
  - low warp
  - high strength

- Dataset:
  - predictors are recipe variables of 4 components
  - 14 response variables are related to different measurements of warp and strength

	predictors
1	glass
2	crtp
3	mica
4	amtp

	response
1	wrp1
2	wrp2
3	wrp3
4	wrp4
5	wrp5
6	wrp6
7	st1
8	st2
9	wrp7
10	st3
11	st4
12	wrp8
13	st5
14	st6

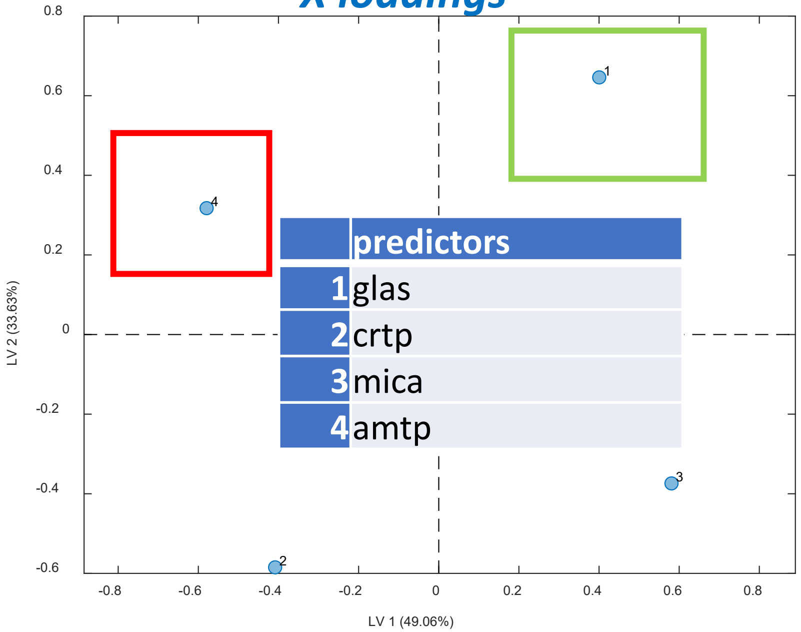
Variables relation

- **X** loading:
  - the inputs are varied with an orthogonal variability through DoE
- **Y** loading:
  - all the strength variables are correlated
  - all the warp variables are correlated
  - warp and strength are anti-correlated on PC1

- Crossed reading:
  - Glass is related to high strength
  - Amtp is related to high warping
  - Crtp is anti-correlated to glass (and high strength)
  - Mica is anti-correlated to amtp (and high warping)

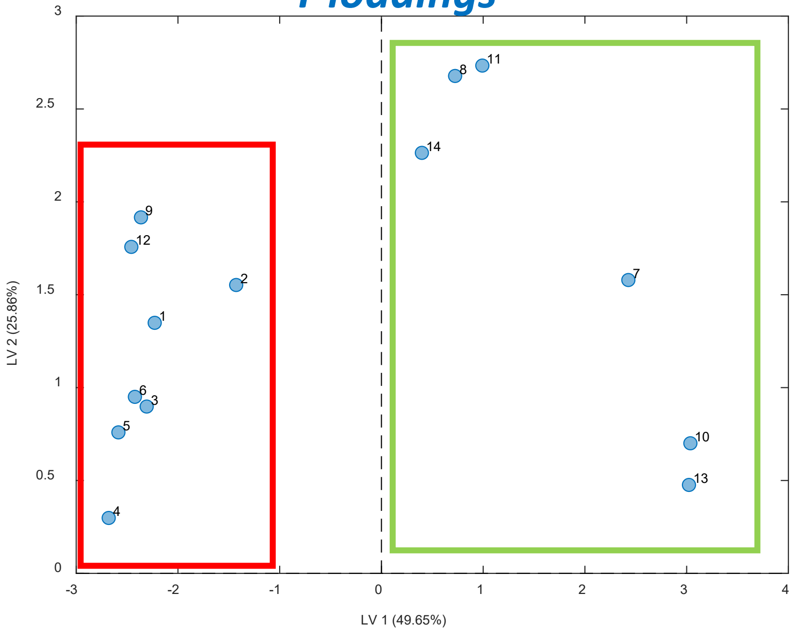
SOLUTION:
1. to increase strength:
   - increase glass
   - lower crtp
2. to decrease warp:
   - increase mica
   - decrease amtp

X loadings



	predictors
1	glas
2	crtp
3	mica
4	amtp

Y loadings



	response
1	wrp1
2	wrp2
3	wrp3
4	wrp4
5	wrp5
6	wrp6
7	st1
8	st2
9	wrp7
10	st3
11	st4
12	wrp8
13	st5
14	st6

Example #2: predictive analysis

exploratory
- understanding correlation within datasets
- understanding relations between datasets

predictive
- **response estimation**
- class attribution in PLS Discriminant Analysis

prescriptive
- design of experiments
- PLS model inversion

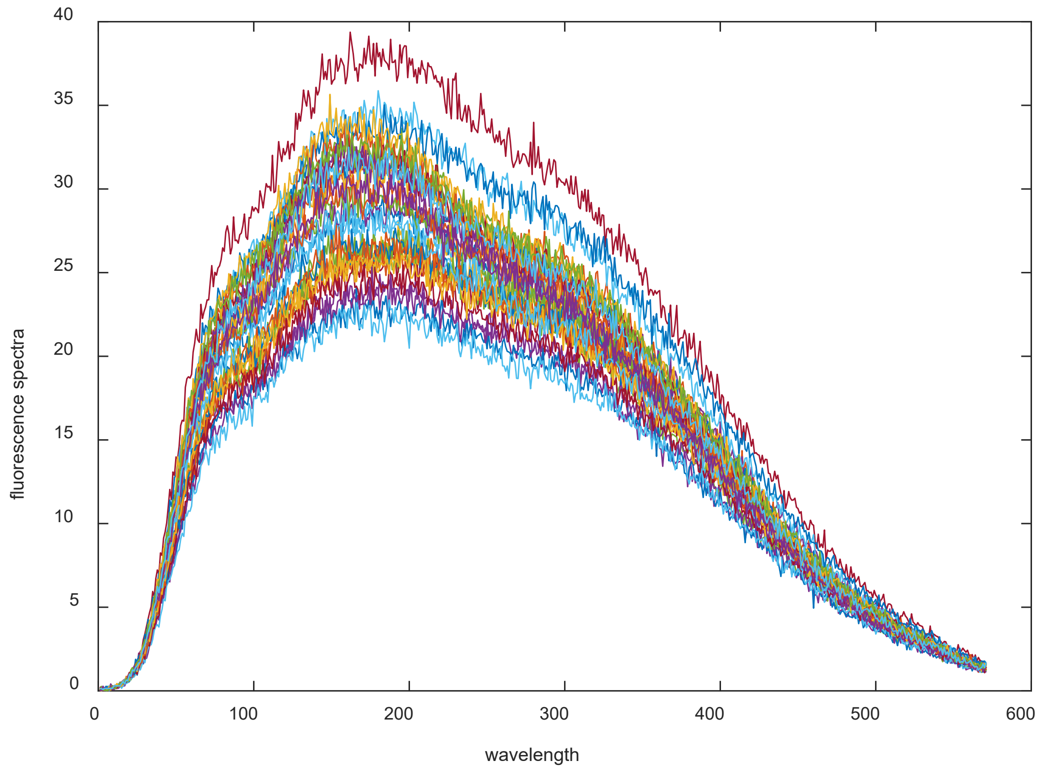
Example of ash content in sugar

Fluorescence spectra

- Industrial production of white sugar

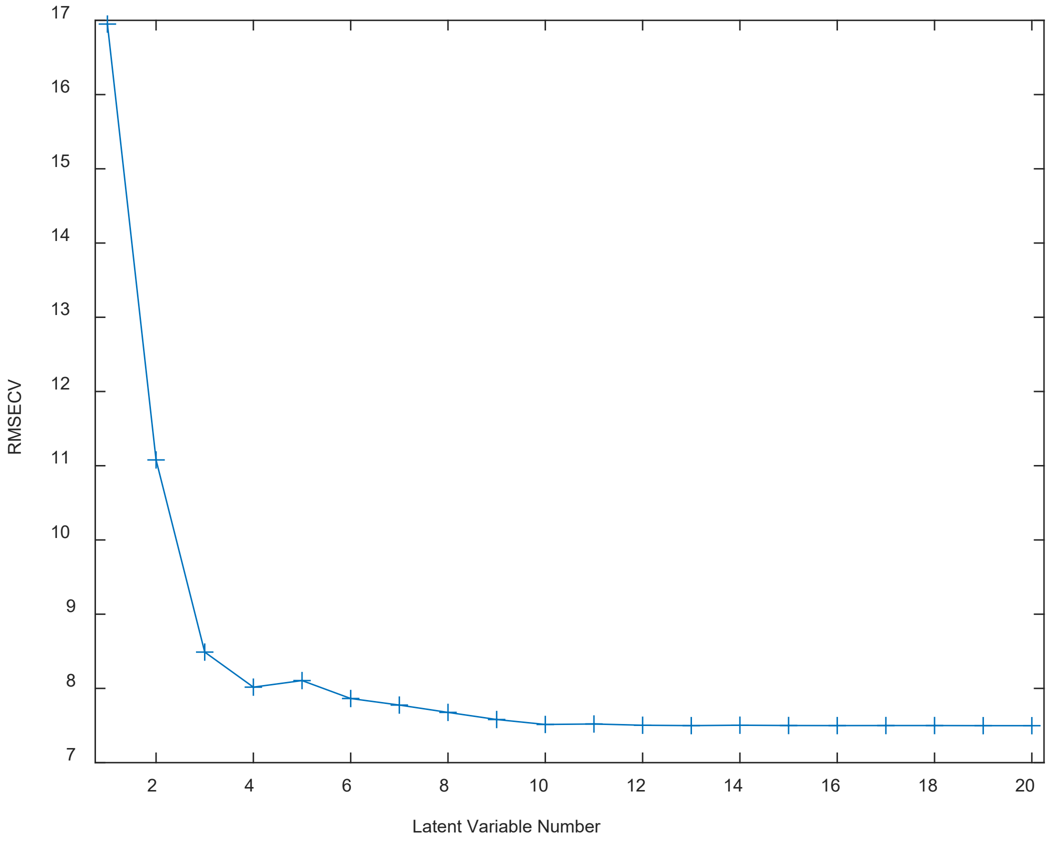
- <u>AIM</u>: estimating the ash content from fluorescence measurements to replace time consuming and expensive chemical measurements

- Available data:
  - predictors:
    - 106 spectra on 500+ wavelengths
    - 53 for validation
  - response:
    - ash content of the sugar
      - chemical measurement



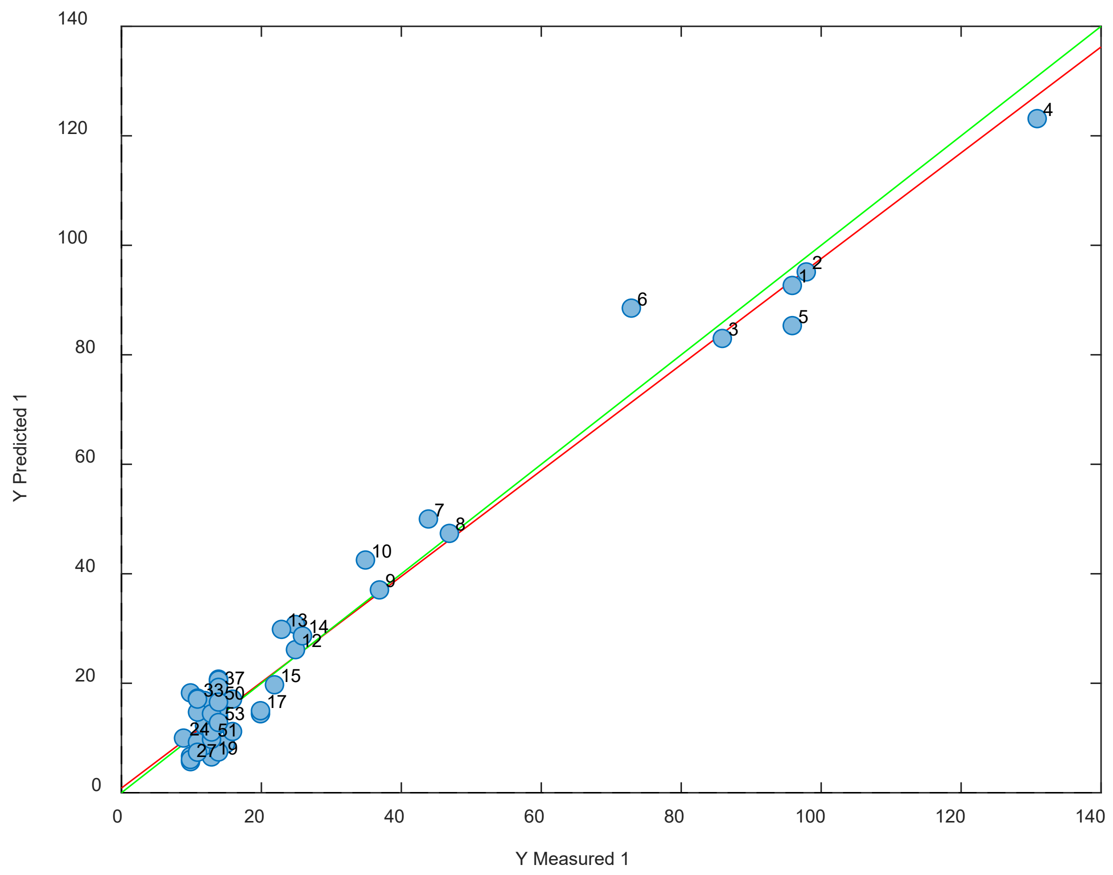
PLS model

- 3 LVs explain 97% of the variability of $\mathbf{Y}$ with 95% of the variability of $\mathbf{X}$



LV	X explained cumulative variance	Y explained cumulative variance
1	78.37	66.49
2	95.57	86.44
3	95.91	96.7

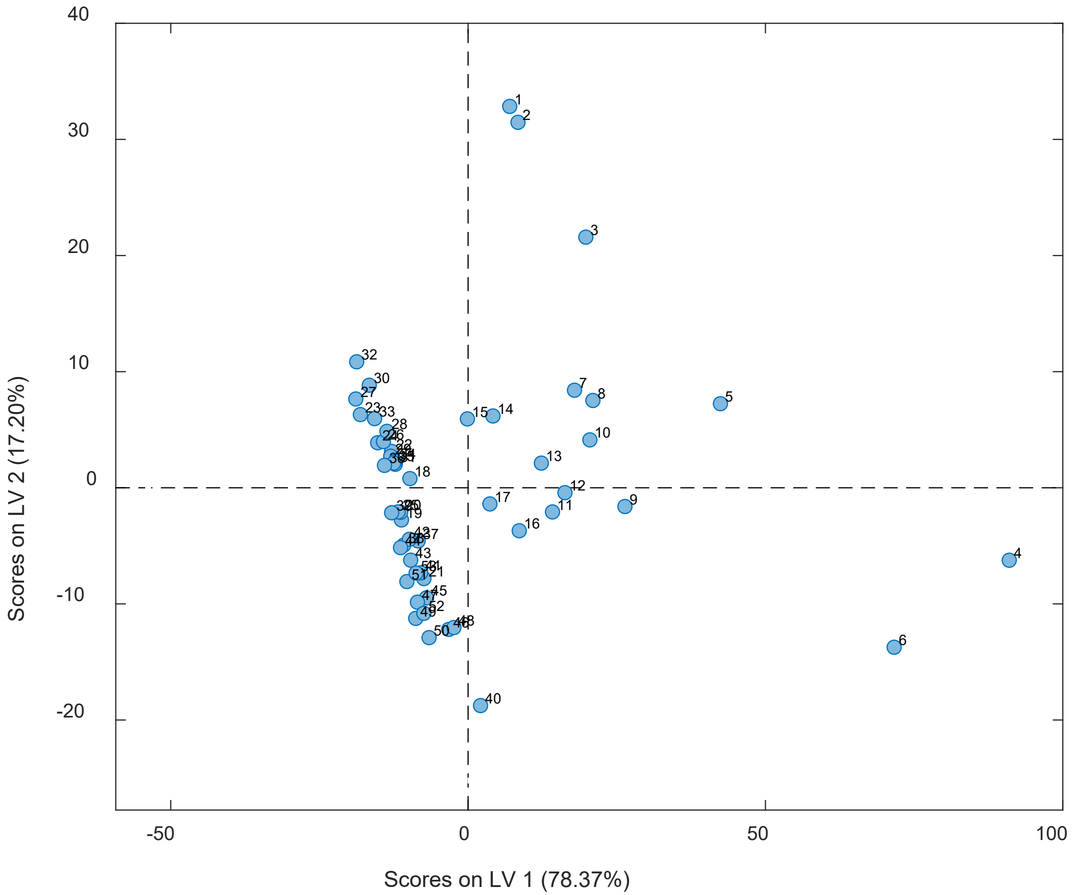
Prediction performance



- Very **promising predictive performance** in calibration
  - parity plot measured vs. predicted is used to access this information

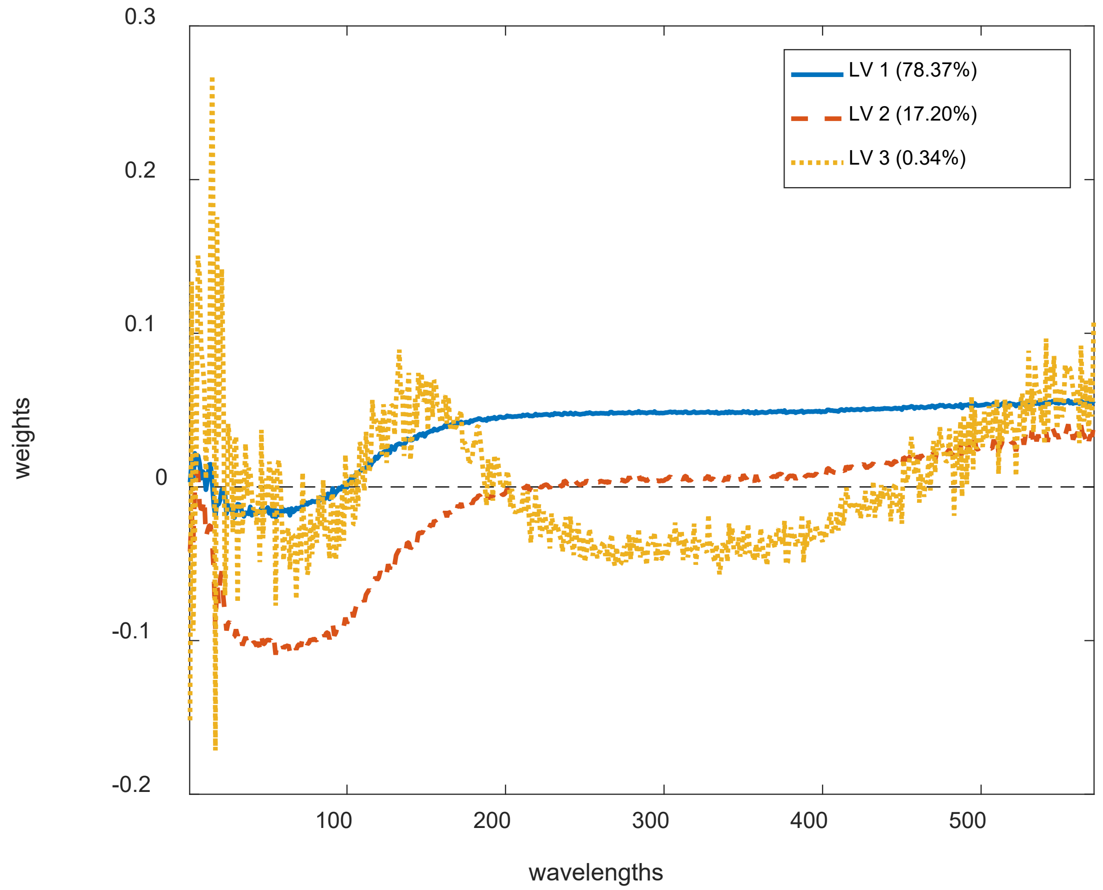
Score plot

- A **process drift** can be seen in the score plot
  - the process variability do not show multivariate Gaussian distribution



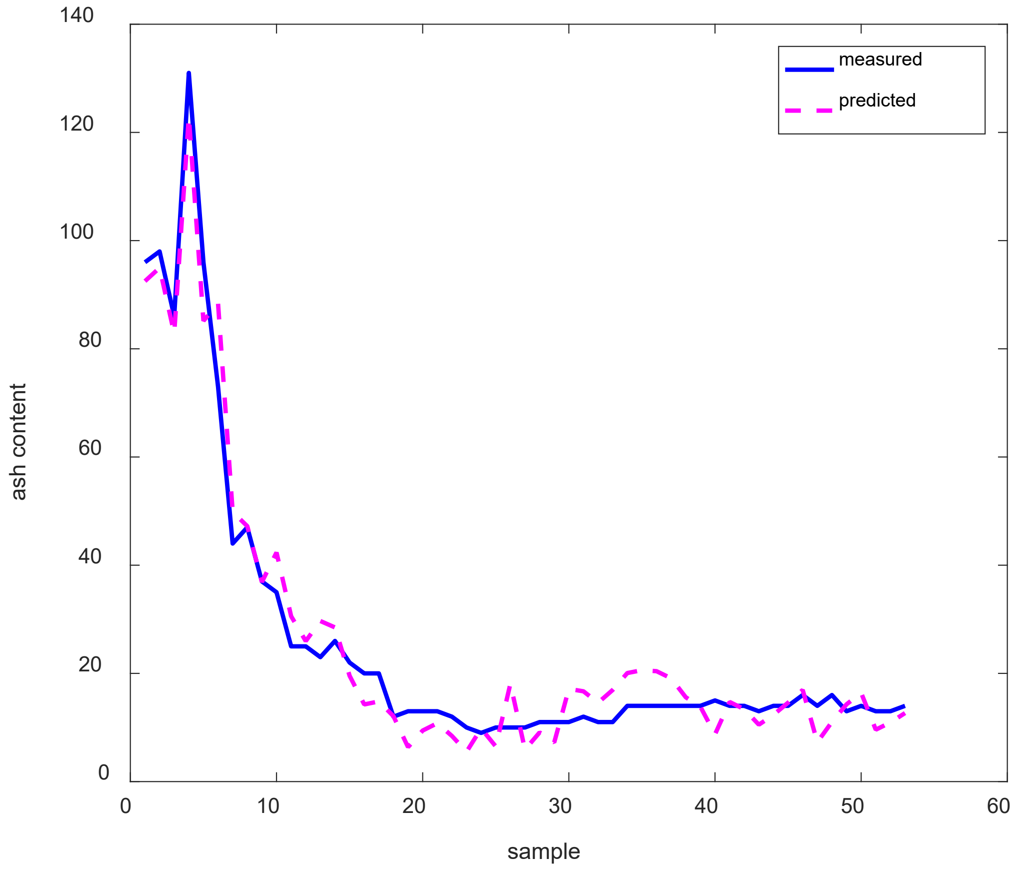
Model weights

- Weights are studied for model interpretation:
  - some wavelengths are more related to the ash content



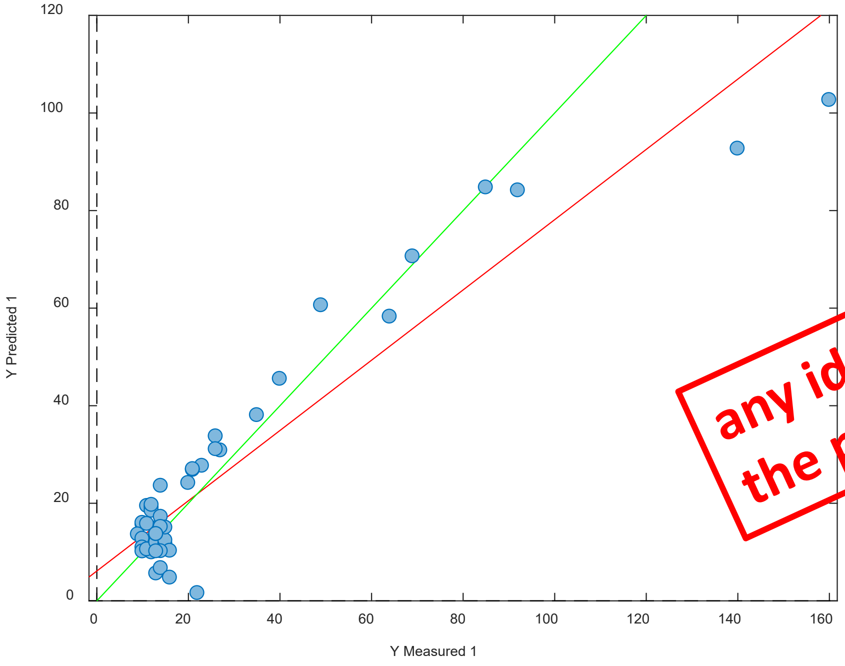
Process drift

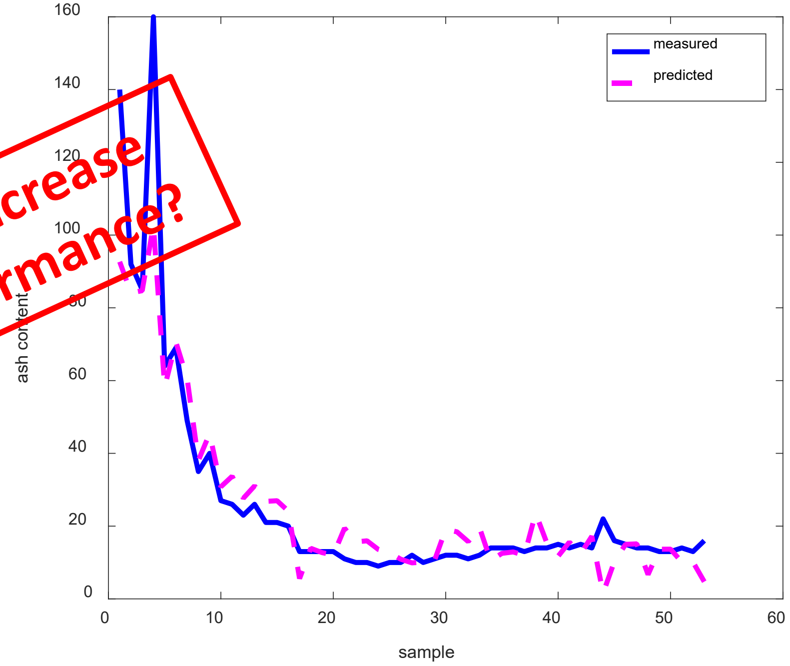
- The ash measurements **Y** are not normally distributed
  - however, the estimation performance are very good:
    - errors are really low
      - maximum error: 15.36
      - average error/st.d.(**Y**) = 0.1448



Predictions for new samples

- Validation: 53 new samples are available
- The prediction performance are slightly lower (as usual…), but acceptable:
  - R^2 decreases to 85.4%
  - the worst performance are present where there is non-linearity and extreme values





any idea on how to increase the prediction performance?

Example #3: classification

exploratory
- understanding correlation within datasets
- understanding relations between datasets

predictive
- **response estimation**
- class attribution in PLS Discriminant Analysis

prescriptive
- design of experiments
- PLS model inversion

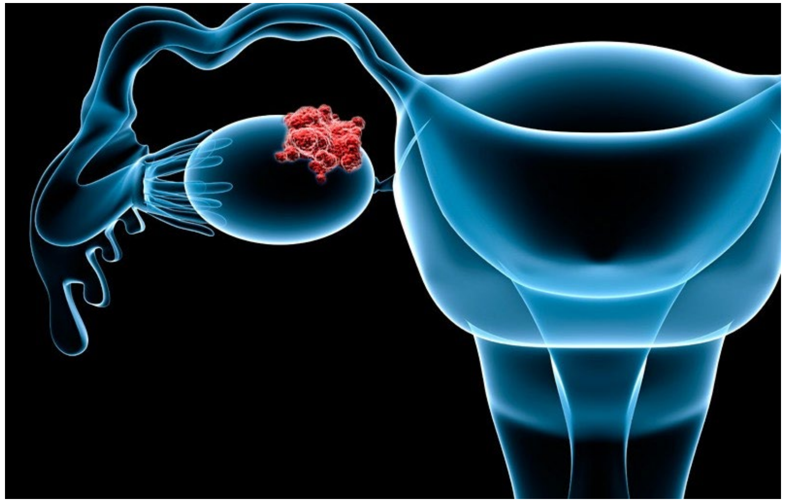
PLS-Discriminant analysis

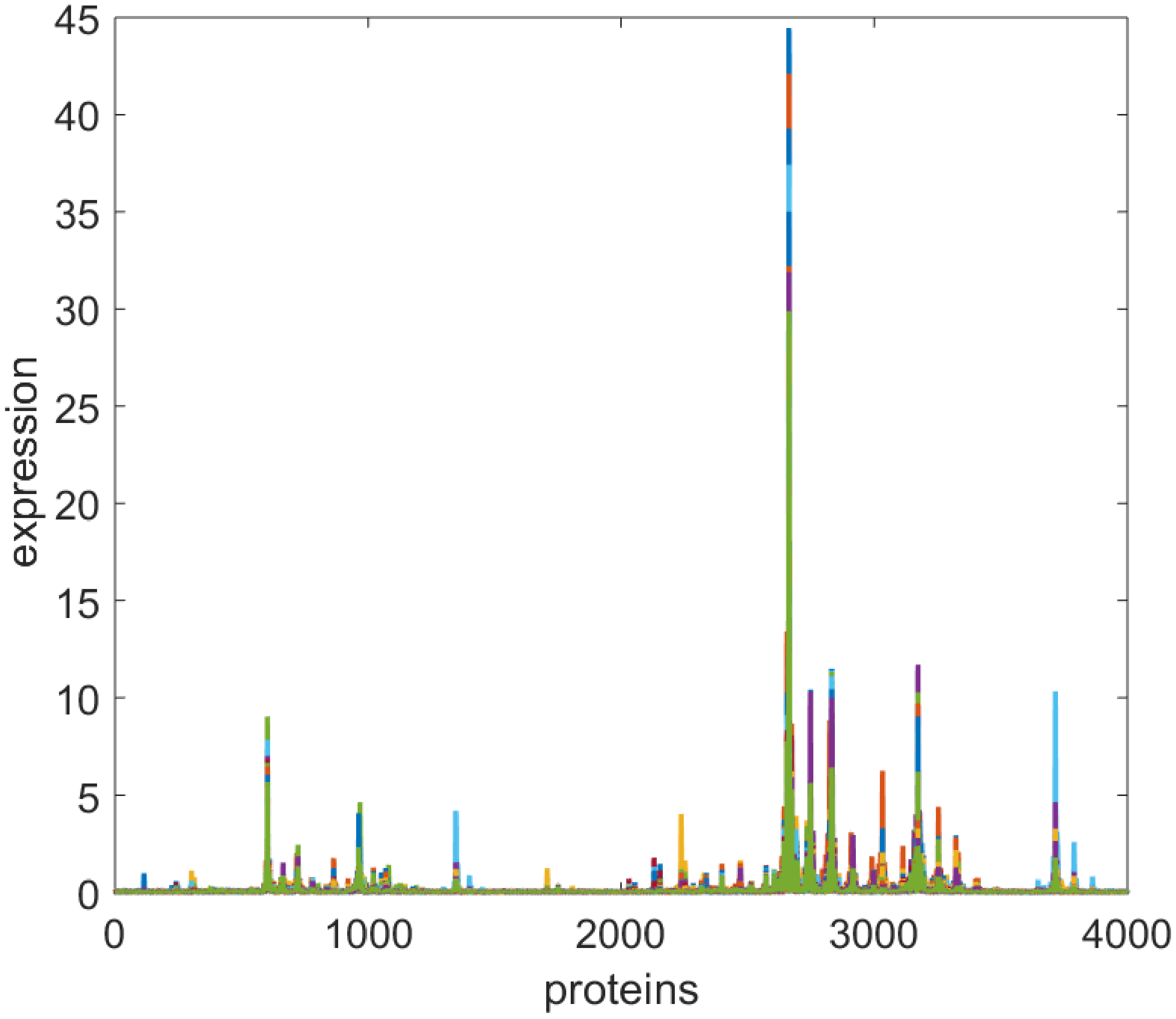
- **PLS-Discriminant Analysis** is the PLS version for classification:
  - the response variable expresses the class
  - the class is expressed as a dummy variable of 0s and 1s in the $\mathbf{Y}$ matrix
    - columns of $\mathbf{Y}$ indicate different classes, rows of $\mathbf{Y}$ indicate different observations
      - $y_{n,m} = 0$ indicates that the observation n does not belong to class m
      - $y_{n,m} = 1$ indicates that the observation n belongs to class m
  - the loadings/weights identify how the classes are characterized by the predictors
  - a probabilistic attribution to the classes is possible, even if the class attribution depends on a threshold between classes

Ovarian cancer classification

Case study

- Dataset on ovarian cancer:
  - 166 calibration observations
    - class1: disease (89 samples)
    - class 2: no disease
  - 50 validation observations
    - class1: disease (32 samples)
    - class 2: no disease
  - 4000 variables: proteomic data

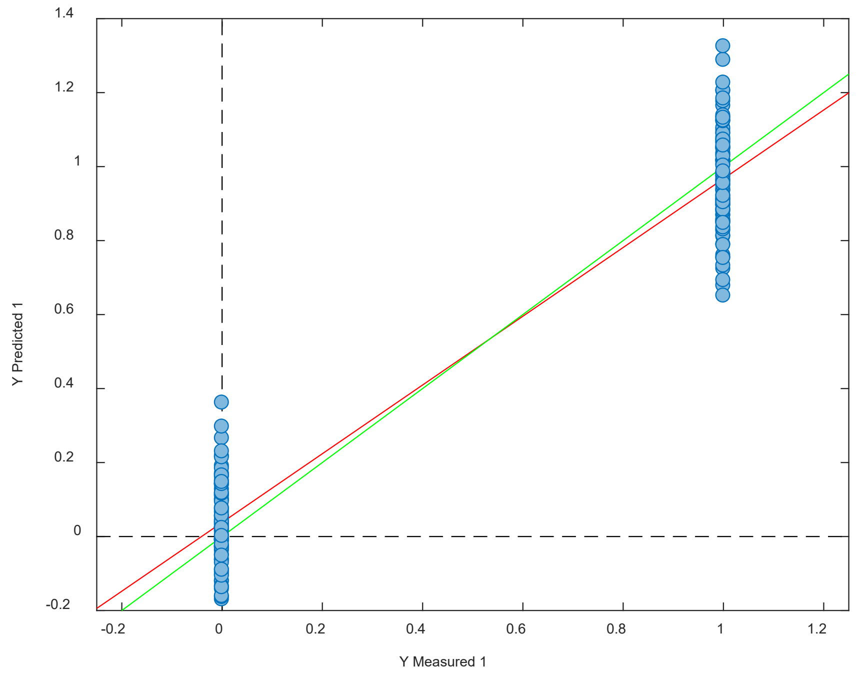




PLS-DA model and discrimination ability

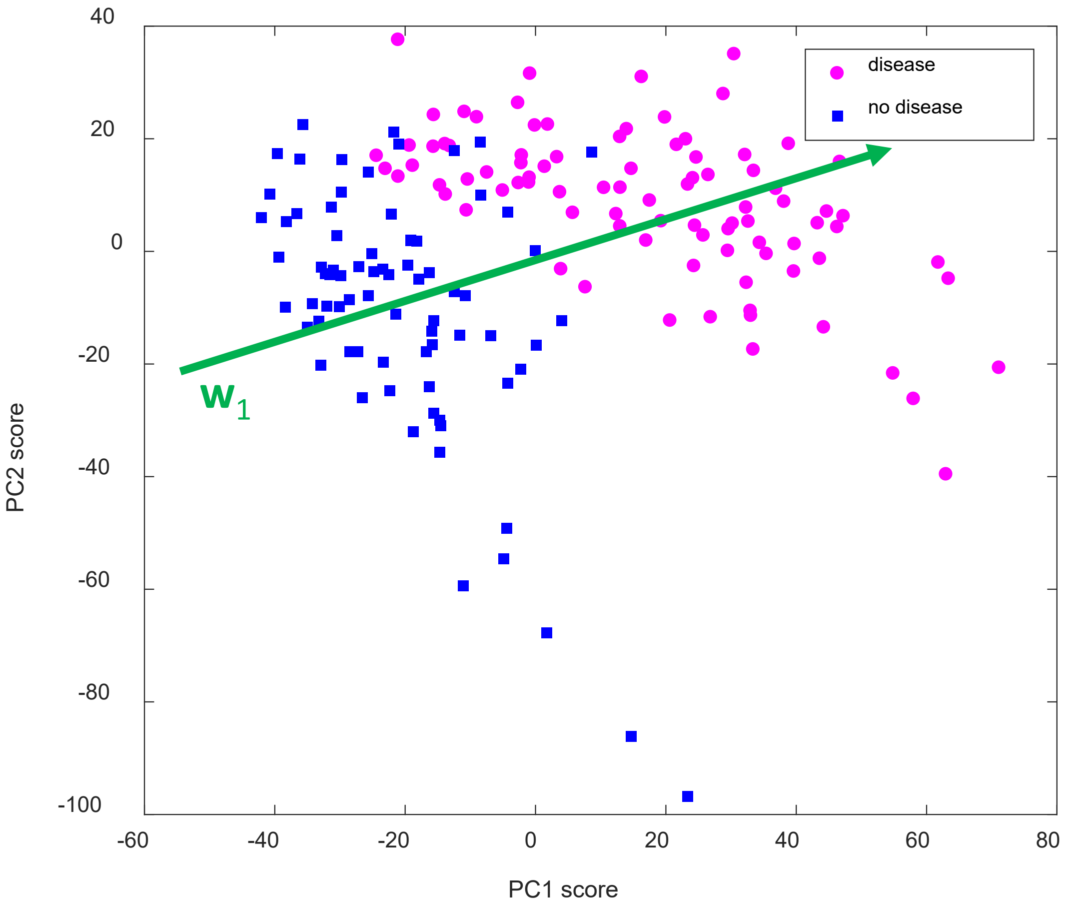
- 4 PC explain > 90% of the class variability with 34% of the predictors' variability
- The **prediction accuracy** (typically identified by means of the percentage of correctly classified observations or the ratio of misclassified observations) in calibration is perfect:
  - 100% of the calibration observations are correctly classified
  - 1 misclassified sample in validation

LV	explained cumulative variance on X	explained cumulative variance on Y
1	18.15	49.16
2	28.71	70.13
3	31.59	87.31
4	34.03	92.91

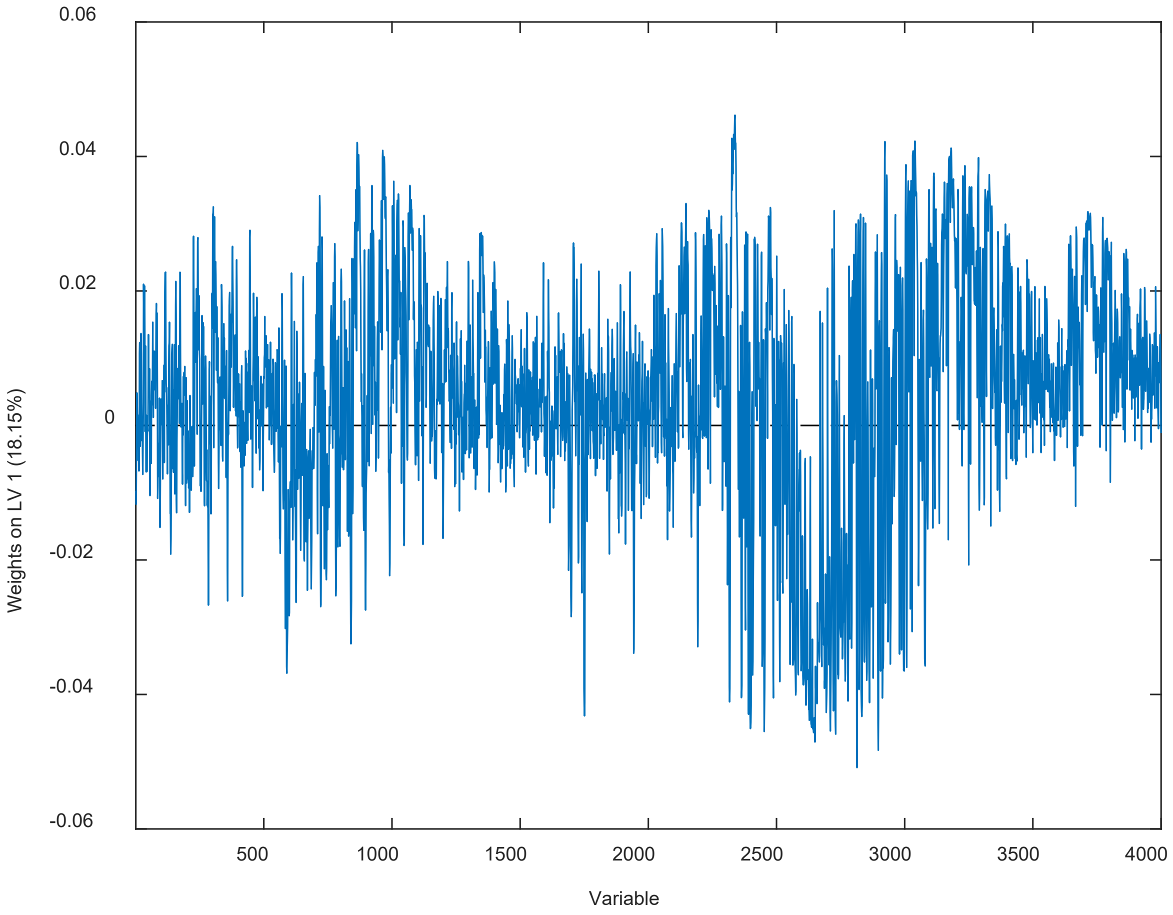


Score space observation discrimination

- Typically, the **score plot** separates the **clusters** of the classes along the direction indicated by the weights of the principal components that are most related to class separation
  - this direction is almost parallel to PC!



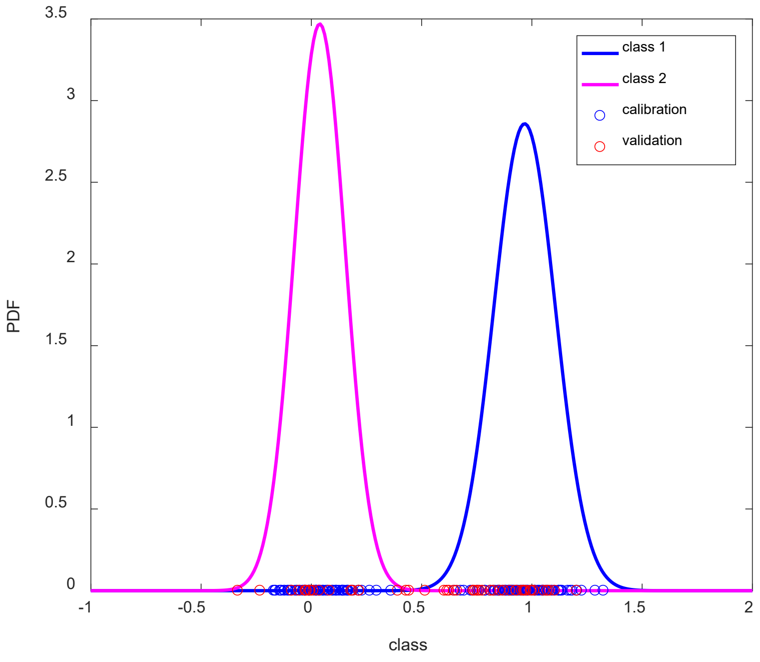
PLS weights and class attribution

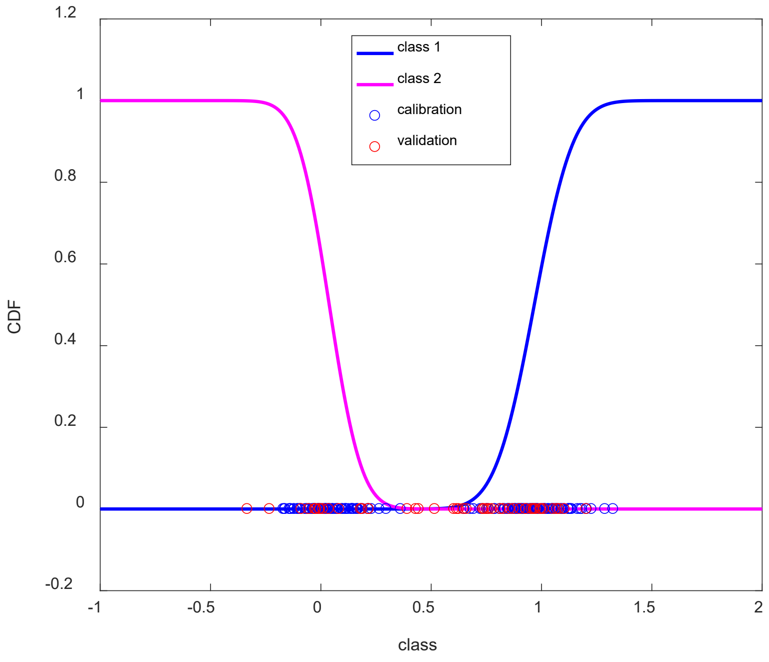


- PLS **weights** are used to understand the predictors (i.e., the proteins) which are more correlated to the class discrimination

Class attribution and probabilistic interpretation

- **Probabilistic attribution of samples** to the class:
  - the attribution to one of the classes is associated with a determined probability of belonging to that class
- Appropriate probabilistic determination of the **threshold** among classes:
  - the threshold is set in the intersection of the cumulative density functions calculated from the predicted values $\hat{y}$ of the calibration observations which are present in each class





Self assessment

- What is PLS?
- What are analogies and differences between PCA and PLS?
- How could you assess information on the correlation among $\mathbf{X}$ and $\mathbf{Y}$?
- What are analogies and differences between PLS and MLR?
- What are analogies and differences between PLS and PCR?
- What are the typical uses of PLS?
- What are the most common applications of PLS?

… per sempre a fianco a me!