Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #10

Academic year 2025-2026

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

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

Recap of the previous lectures

- We introduced **latent-variables projection methods** to deal with multivariate correlated data
- **PCA** is the basis of the multivariate statistical methodologies
  - utilized for:
    - data dimensionality reduction
    - process understanding and troubleshooting
  - easy to identify:
    - relationships between observations
    - correlation between variables
- **SPC** provides valuable tools for process and product quality monitoring
  - control charts
    - univariate: Shewhart chart
    - multivariate?

Today's lesson

- Revise univariate monitoring
- Multivariate monitoring
  - multivariate control charts
  - multivariate control limits

Examples on monitoring applications

Univariate SPC on bolt size

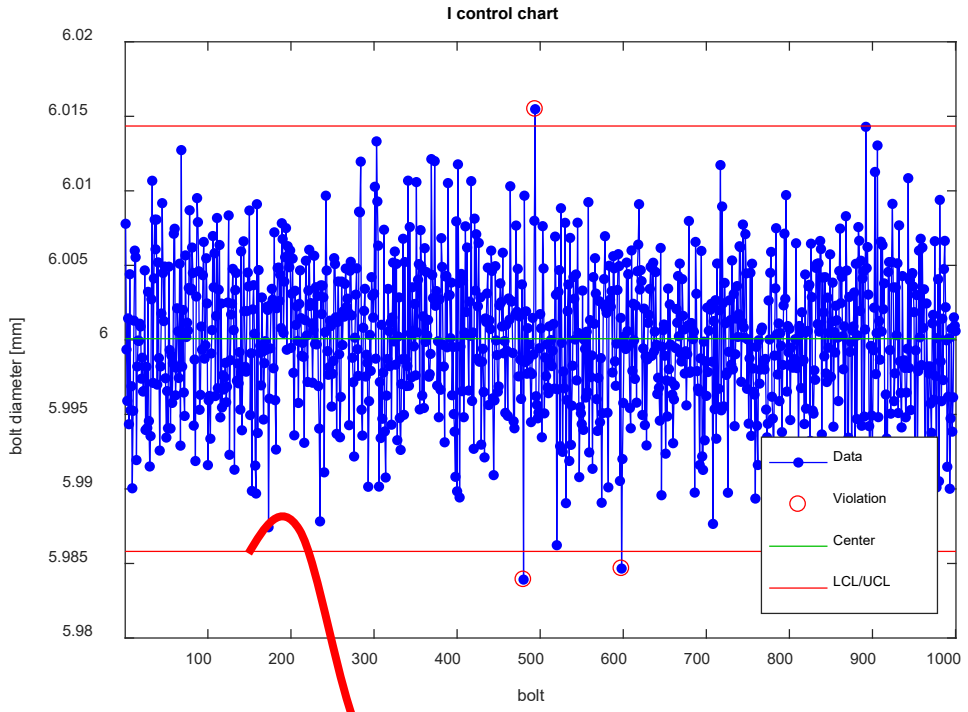
Monitoring the diameter of the bots

- Consider the case of bolt production
- The “natural variability” that affects the process of bolts production allows building univariate SPC monitoring charts on the product quality
  - the variability defines the confidence limits

Univariate monitoring of the bolt diameter

- Example:
  a **Shewhart chart** is built on 1000 observations

- What do you expect?
  - the sample mean is close to the desired value of 6 mm
  - the sample standard deviation is close to 0.005 mm
  - 3 samples over 1000 are (by definition) out of 3σ limits



the Shewhart chart confidence limits in Matlab® are at 3σ

```
controlchart(X(:,1),'chart',{'i'})
```

Univariate and multivariate SPC on athletes’ shape

Monitoring athlete's height and weight

- Consider the problem of athletes' weight and height as a monitoring problem

athlete	height (cm)	weight (kg)
1	175	73
2	198	110
3	168	65
4	182	95
5	178	81
6	185	99
7	177	80
8	171	105
9	188	83
10	176	74

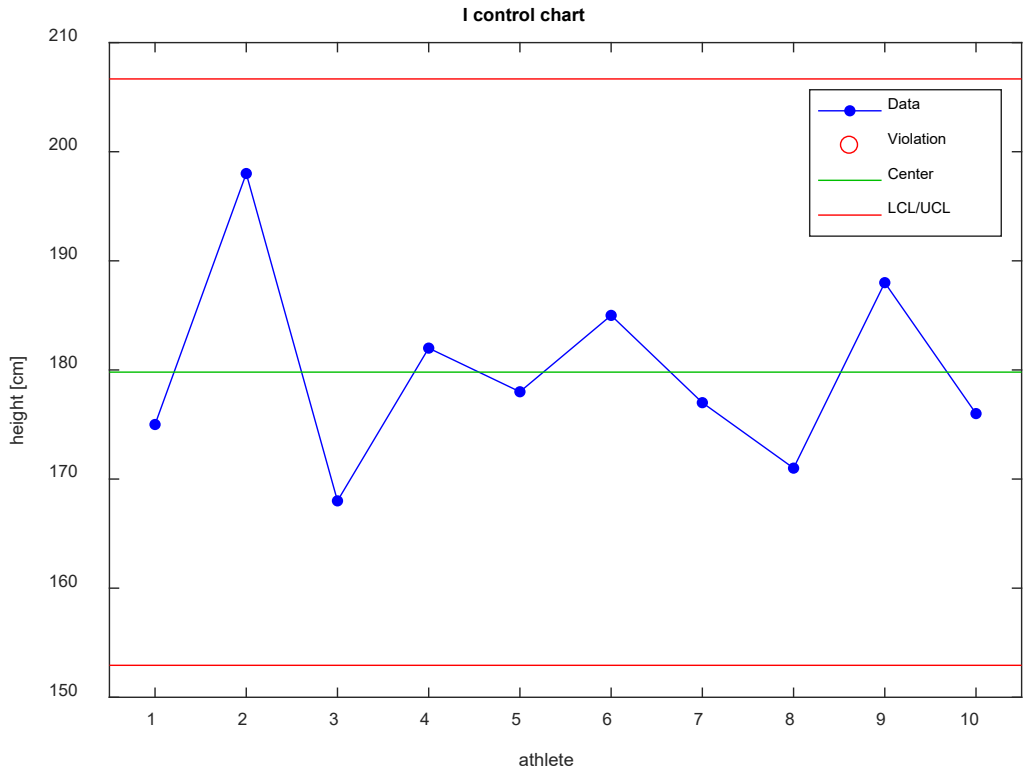
Univariate monitoring of the height

- The **Shewhart chart** fixes the target and the normal variability

$$\bar{x} = 179.8 \text{ cm}; \quad S = 8.8 \text{ cm}$$

  - no “special” variability is identified

athlete	height (cm)
1	175
2	198
3	168
4	182
5	178
6	185
7	177
8	171
9	188
10	176

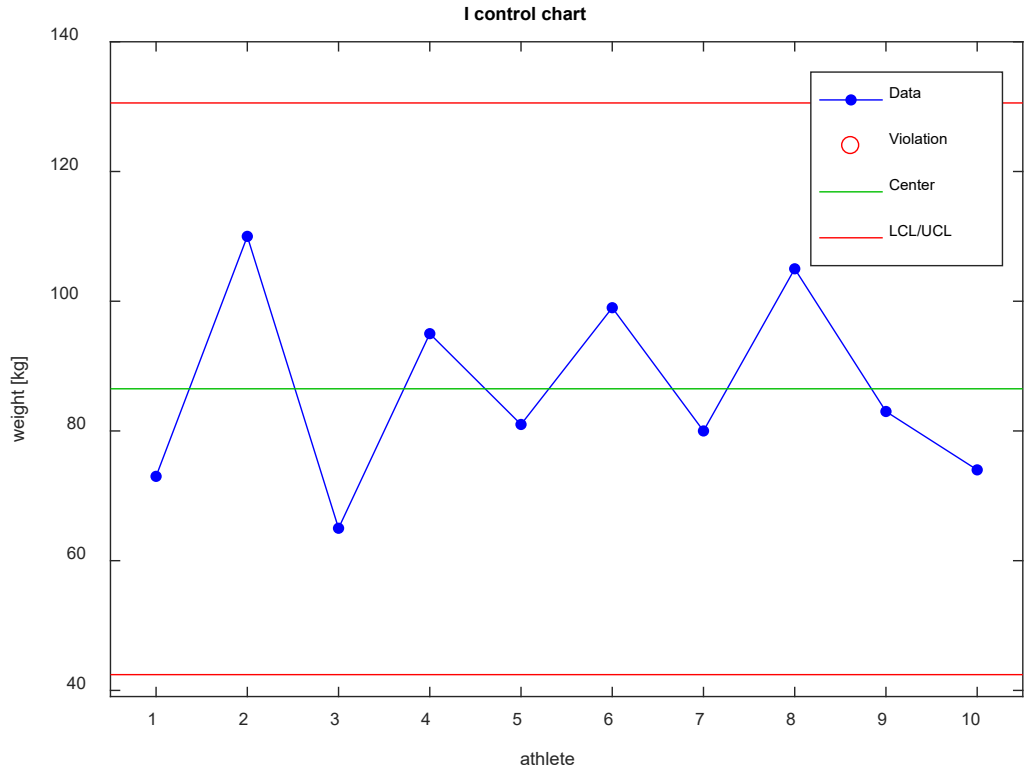


Univariate monitoring of the weight

- From the Shewhart chart on the body weight no anomalies are detected

$$\bar{x} = 86.5 \text{ kg}; \quad S = 14.9 \text{ kg}$$

athlete	weight (kg)
1	73
2	110
3	65
4	95
5	81
6	99
7	80
8	105
9	83
10	74

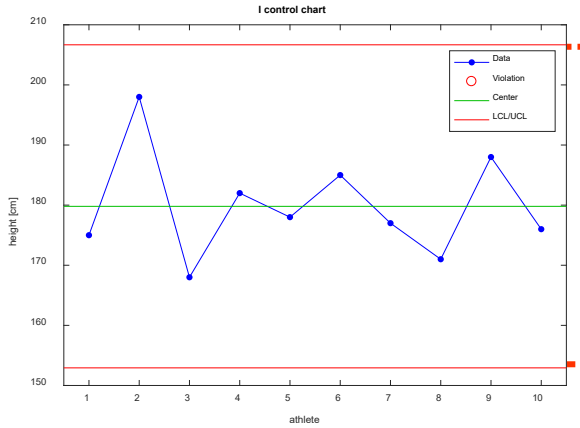


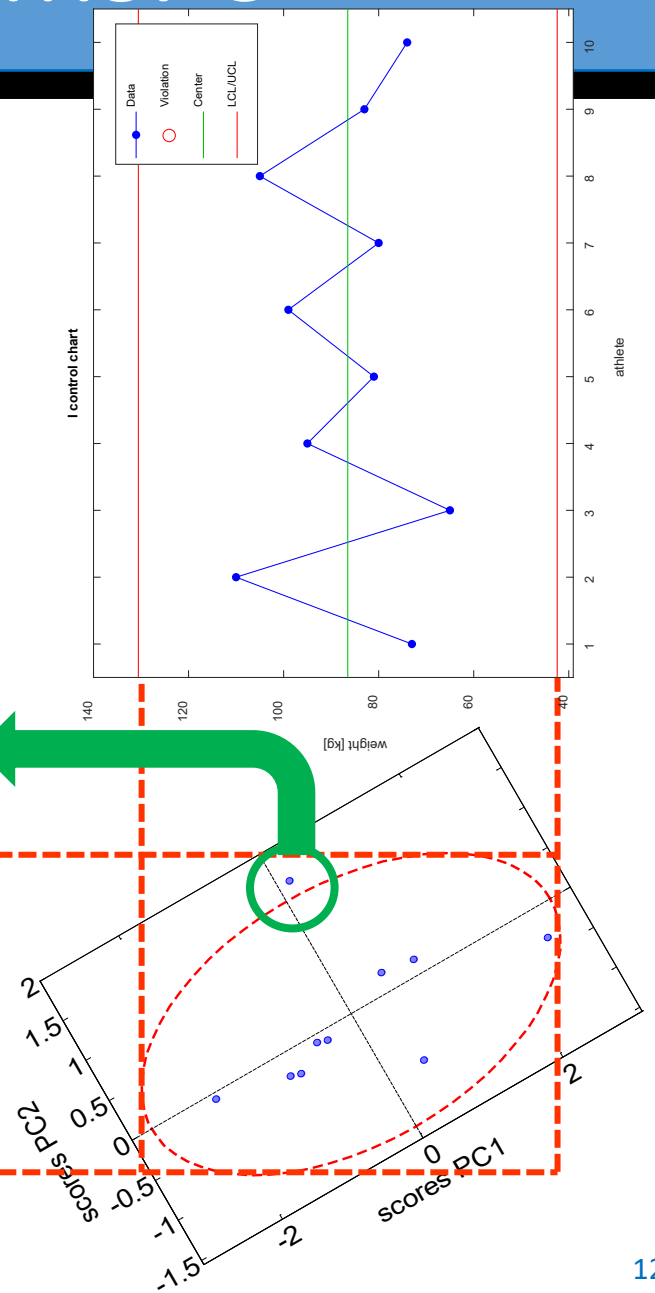
Multivariate statistical process control MSPC

- The **univariate control charts cannot detect any deviating behavior related to data correlation**
- **Multivariate control charts** consider correlation among variables:
  - **multivariate confidence limits** in a multivariate chart are:
    - elliptical in the score space
    - not simply the union of the Shewhart-charts confidence limits

athlete	height (cm)	weight (kg)
1	175	73
2	198	110
3	168	65
4	182	95
5	178	81
6	185	99
7	177	80
8	171	105
9	188	83
10	176	74

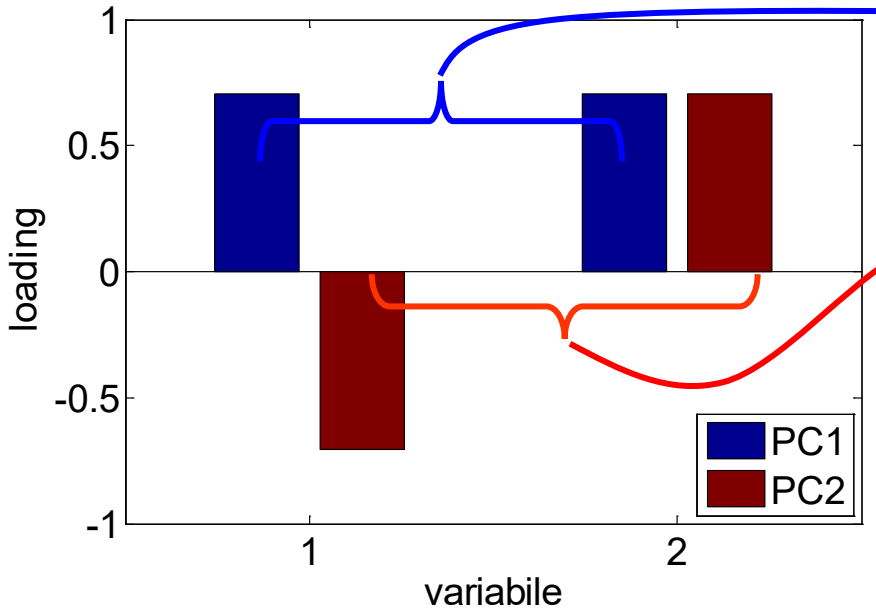
the anomaly is evident in a MULTIVARIATE control chart!!!





PCA monitoring model on body shape (1/2)

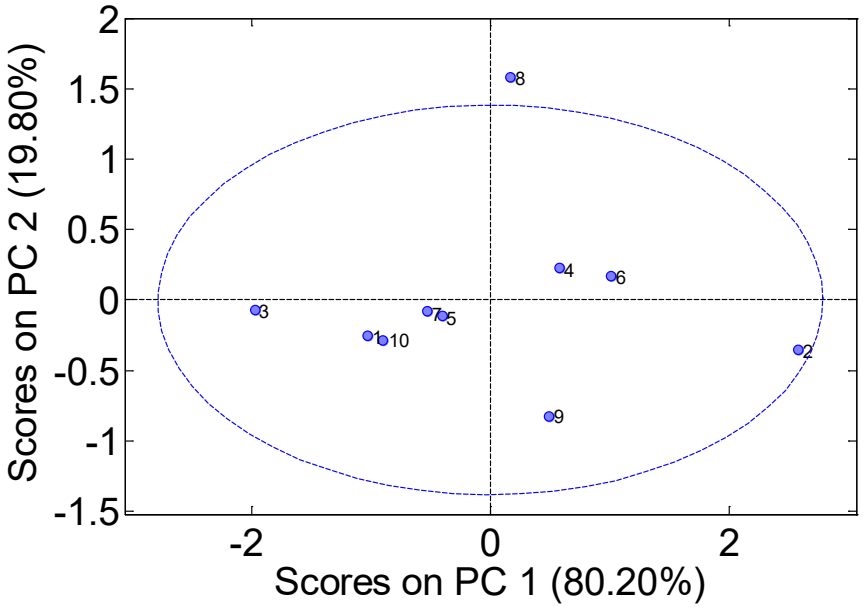
loading plot – variables correlation



80% of data variability explains the **positive linear correlation** between height and weight

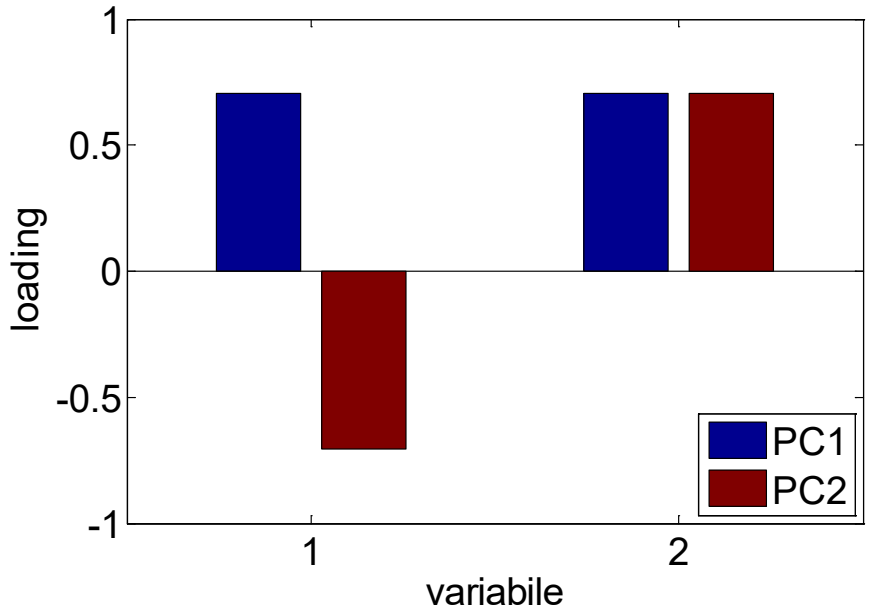
20% of data variability explains the **anti-correlation** between height and weight

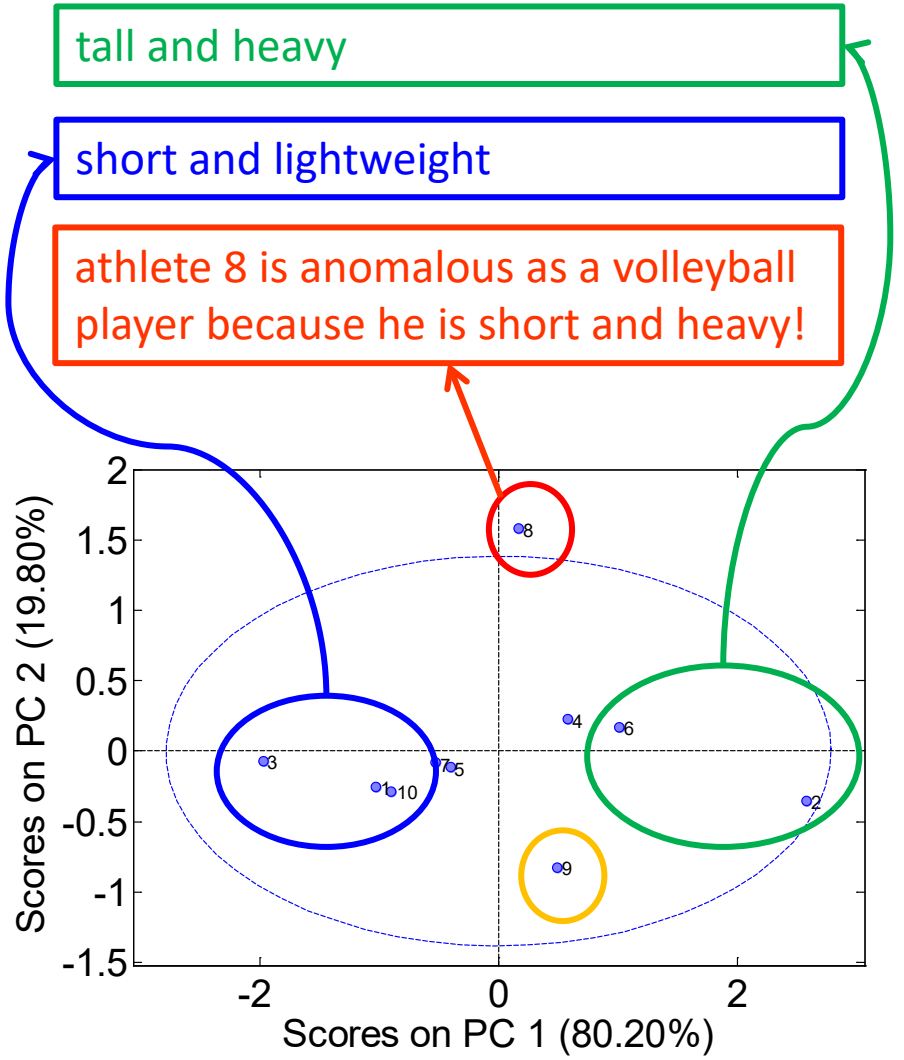
score plot – relations among observations



PCA monitoring model on body shape (2/2)

athlete	height (cm)	weight (kg)
1	175	73
2	198	110
3	168	65
4	182	95
5	178	81
6	185	99
7	177	80
8	171	105
9	188	83
10	176	74



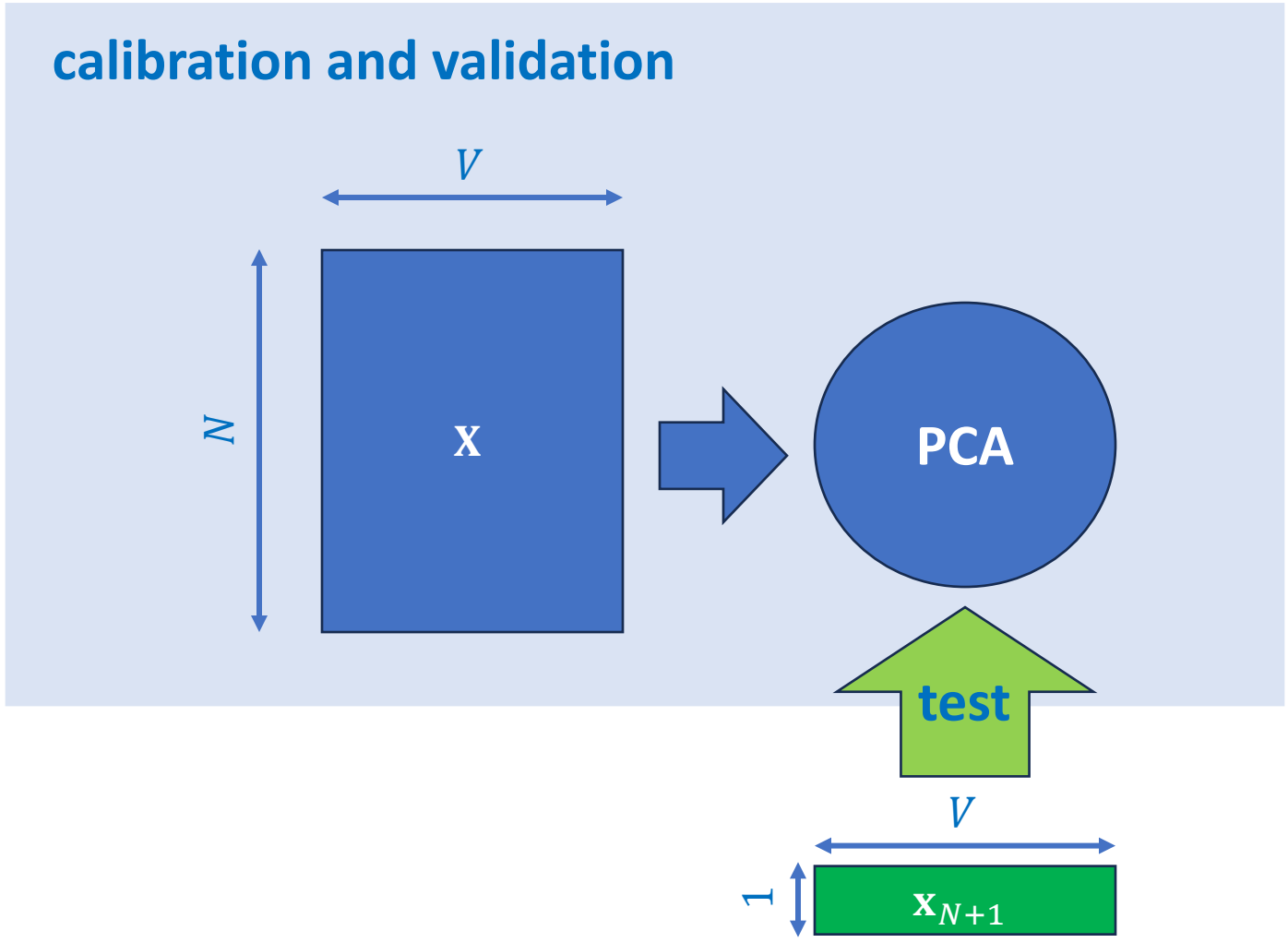


Calibration, validation and testing of PCA monitoring

- PCA modelling is very effective for exploratory analysis
  - “passive” analysis on data relations
- However, PCA can be utilized in a more **proactive** way: **multivariate monitoring**

- Phases of PCA modelling for monitoring:
  - **calibration**
    - model building on a selected dataset on normal operating (standard) conditions of the process/system (product) under study
  - **validation**
    - projection of new unknown data into the space of principal component previously calibrated
  - **testing**
    - application to real-life!

Schematic of the use of PCA for SPC



Dealing with new observations

- Once the model is calibrated, the conformance of a **new observation** $\mathbf{x}_{N+1}\ [1 \times V]$ to the calibration reference can be checked through the investigation of the abovementioned parameters (projecting it onto the PCA model):
  - to assess the *deviation of the new observation from the average conditions of the reference*
  - to evaluate the model representativeness, namely *how well the model fits the actual conditions of the new incoming observation*

- These concepts can be (easily) translated in statistical indices, which are synthetic responses on the status of the new observation, and are used to build **multivariate statistical monitoring charts**

Group discussion

- What are the statistical test involved in these control charts?
  - what are the statistical indices you would use for multivariate monitoring charts?
- How would you build a control chart on these statistical indices?
  - you already know it… :)

Statistical tests on PCA sample diagnostics

- **Statistical tests** are established to evaluate if, based on the T^2 and SPE statistics a sample should be considered, with a predetermined degree of confidence:
  - **regular**: it is inside the confidence region of the standard observations
  - **irregular**: it is in the rejection region of the observations that are not considered to be standard
- The pieces of information provided by the T^2 and SPE statistics are:
  - observations with **high values of T^2** are characterized by the same correlation structure as the one described by the PCA model, but extreme values of some of the most important variables (variables most impacting the variability)
    - in calibration these observations has a strong influence (i.e., **leverage**) on the model
      - in determined conditions they must be removed
  - observations with **high values of SPE** are characterized by a different correlation structure compared to the one described by the PCA model
    - in calibration samples with high values of SPE, but low values of T^2, do not have much influence on the model and do not provide information
      - by removing them the model is unlikely to change

PCA monitoring charts application

PROCEDURE: when a **new observation** $\mathbf{x}_{N+1}$ $[1 \times V]$ is available:

1. **project it into the model built on the calibration dataset** (which are the selected reference NOC observations):

$$\hat{\mathbf{t}}_{N+1} = \mathbf{x}_{N+1}\mathbf{P}$$

2. **calculate the sample diagnostics** Hotelling T^2_{N+1} and squared prediction error Q_{N+1} for the new observation:

$$T^2_{N+1} = \hat{\mathbf{t}}^{\mathrm{T}}_{N+1}\Lambda^{-1}\hat{\mathbf{t}}_{N+1} = \sum_{a=1}^{A}\frac{\hat{t}^2_{a,N+1}}{\lambda_a}$$

$$Q_{N+1} = \mathbf{e}^{\mathrm{T}}_{N+1}\mathbf{e}_{N+1}$$

3. compare the statistical indices with the respective **confidence limits** T^2_{lim} and Q_{lim} at a predetermined confidence level $100(1-\alpha)\%$

4. two different scenarios can be found:
   - $T^2_{N+1} \leq T^2_{\mathrm{lim}}$ and $Q_{N+1} \leq Q_{\mathrm{lim}}$ ⇒ **regular observation**
   - $T^2_{N+1} > T^2_{\mathrm{lim}}$ or $Q_{N+1} > Q_{\mathrm{lim}}$ ⇒ **anomalous observation?**

5. in case of detected anomalies, the cause can be found **diagnosing** the variables that are mostly related to the anomaly:
   - if the anomaly is found on T^2_{N+1}, then calculate the contribution of the v-th variable to the Hotelling statistics:

$$c^{T^2}_{N+1,v} = \hat{\mathbf{t}}_{N+1}\Lambda^{-\frac{1}{2}}\mathbf{p}^{\mathrm{T}}_{v}$$

   - if the anomaly is found on Q_{N+1}, then calculate the contribution of the v-th variable to the squared prediction error:

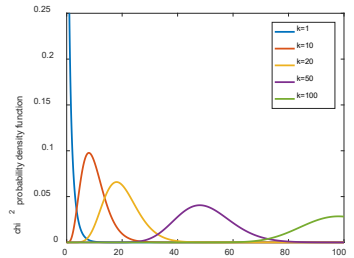
$$c^{Q}_{N+1,v} = e_{N+1,v}$$

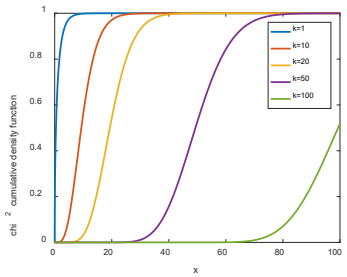
Recalling a little bit of statistics… (1/2)

- The residuals $\mathbf{E}$ of a PCA should retain the non-systematic part of the signal variability (i.e., noise)
  - noise should be a normally-distributed random variable
  - SPE residuals are sum of squares of normally-distributed residuals $\boldsymbol{e}$

- SPE are distributed as a χ^2 distribution

( )





20/03/2025 Prof. Pierantonio Facco - Machine learning for process engineering 6

Recalling a little bit of statistics… (2/2)

- The Hotelling statistics is:

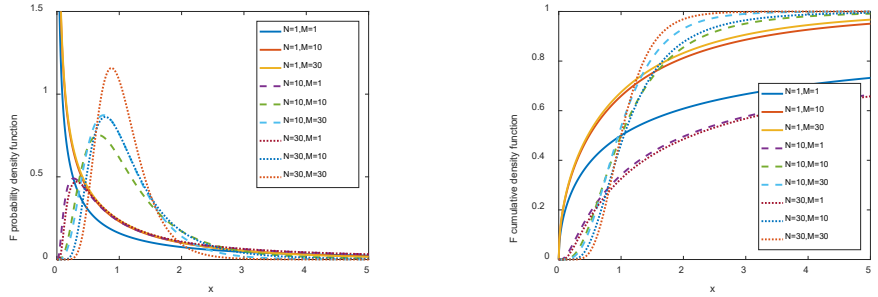
$$T_n^2 = \hat{\mathbf{t}}_n^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \hat{\mathbf{t}}_n = \sum_{a=1}^{A} \frac{\hat{t}_{a,n}^2}{\lambda_a}$$

  - the eigenvalues λ_a are measurements of the variance σ_a^2 explained by the a-th PC
  - T^2 is the ratio between
    - squared normally-distributed scores $\boldsymbol{t}$
    - χ^2-distributed σ_a^2

- T^2 is distributed as a F distribution

Fisher's F distribution and its PDF and CDF

- This distribution is very important in the statistical **analysis of designed experiments** and **statistical process control**:
  - the distribution was presented by the father of the Design of Experiments (DoE)
- Examples of the PDF and the CDF are reported for different values of the parameters N and M



Multivariate confidence limits

- The **Hotelling confidence limit T^2_{lim}** is calculated from a F-distribution with A and $(N-A)$ degrees of freedom for a confidence level $100(1-\alpha)\%$:

$$T^2_{\lim} = \frac{A(N-1)}{N-A} F_{A,N-A,1-\alpha}$$

  - the **F-distribution** is the distribution of the sum of squares of Gaussian observations divided by their variance
  - the semi-axis of the **confidence ellipsoid in the score plot** are:

$$l_a = \sqrt{\lambda_a T^2_{\lim}}$$

- The **squared prediction error confidence limit Q_{lim}** at a confidence level of $100(1-\alpha)\%$ can be calculated as:
  - a **χ^2-distribution**, since it is the sum of squares of errors that should be normally distributed:

$$Q_{\lim} = \frac{s_Q^2}{2\mu_Q} \chi^2_{\frac{2\mu_Q^2}{s_Q^2},1-\alpha}$$

  - alternatively, from the **Jackson-Mudholkar equation**:

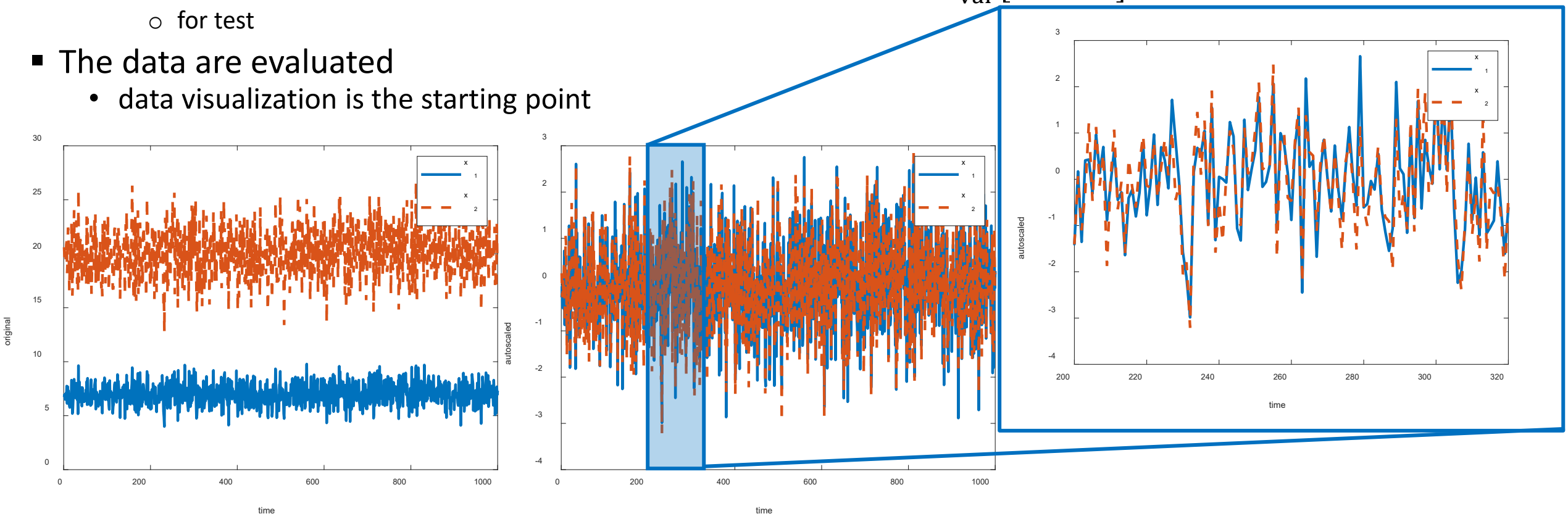
$$Q_{\lim} = \theta_1 \left(\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{\frac{1}{h_0}}$$

$$\theta_i = \sum_{j=A+1}^{R} \lambda_j^i$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

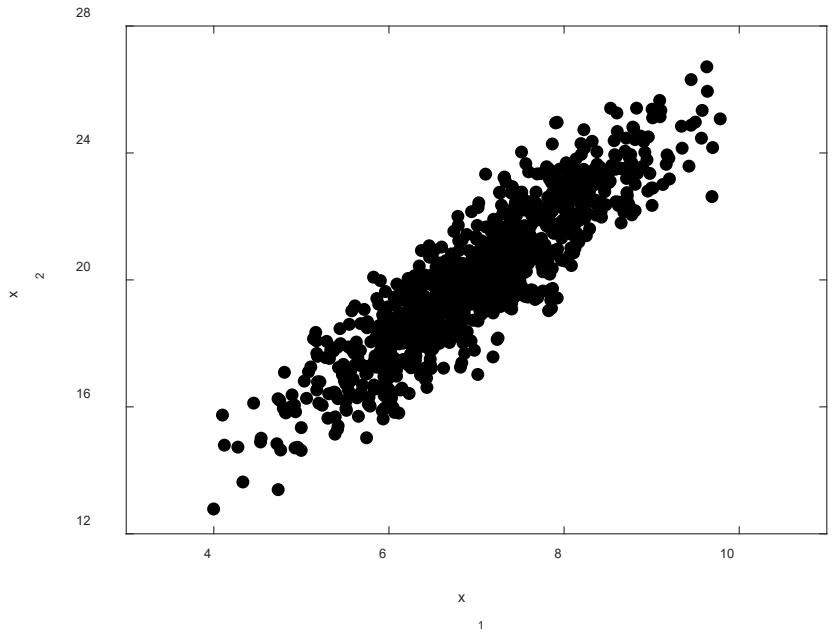
Example of a simple continuous process

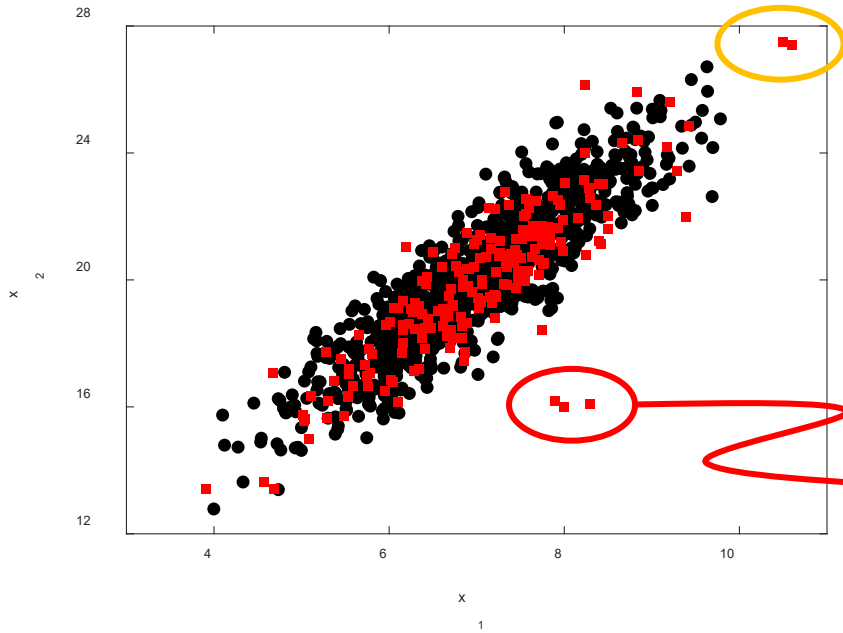
- Process industry case study:
  - 2 process measurements (x_1, x_2) are collected in real time by process computers
  - 1000 observations are considered from the standard operating conditions in matrix $\mathbf{X}_{\text{cal}}$ $[1000 \times 2]$
    - considered for calibration and validatio
  - 205 extra observations of the same measurements are collected $\mathbf{X}_{\text{val}}$ $[205 \times 2]$
    - for test
- The data are evaluated
  - data visualization is the starting point



Variables correlation

- A joint view of the data is missing, so multivariate methods are needed:
  - a common direction of variability is evident
    - where variables co-vary, namely, are strictly correlated $\left(\rho_{x_1,x_2} = +0.895 \sim +1\right)$
- The test observations are projected onto the calibration data:
  - standard data are superimposed to the common direction of variability
  - same data, even if at **standard values**, **deviate from the expected correlation**
  - some data, even if with a **standard correlation**, **deviate from standard values**



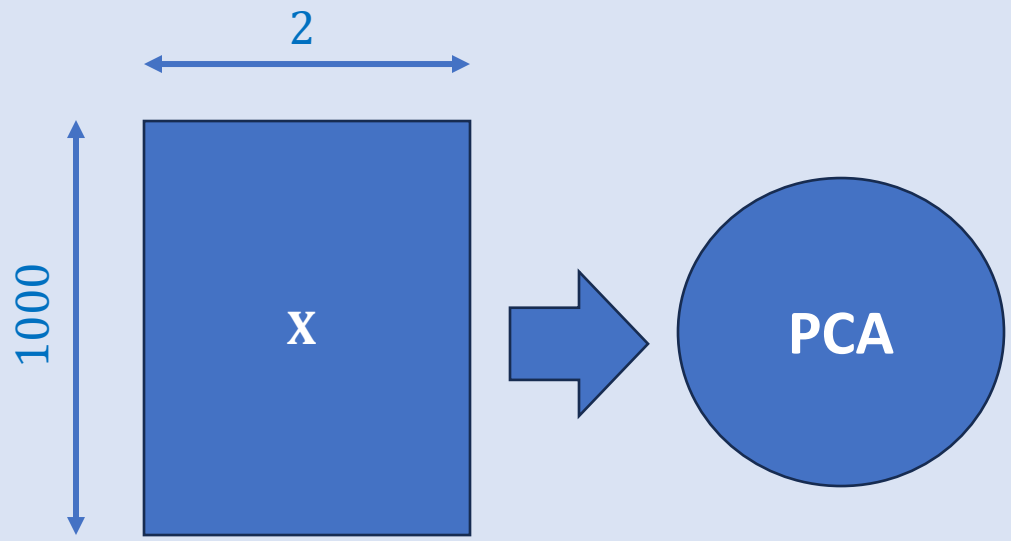


standard correlation, but anomalous values

standard values, but broken correlation

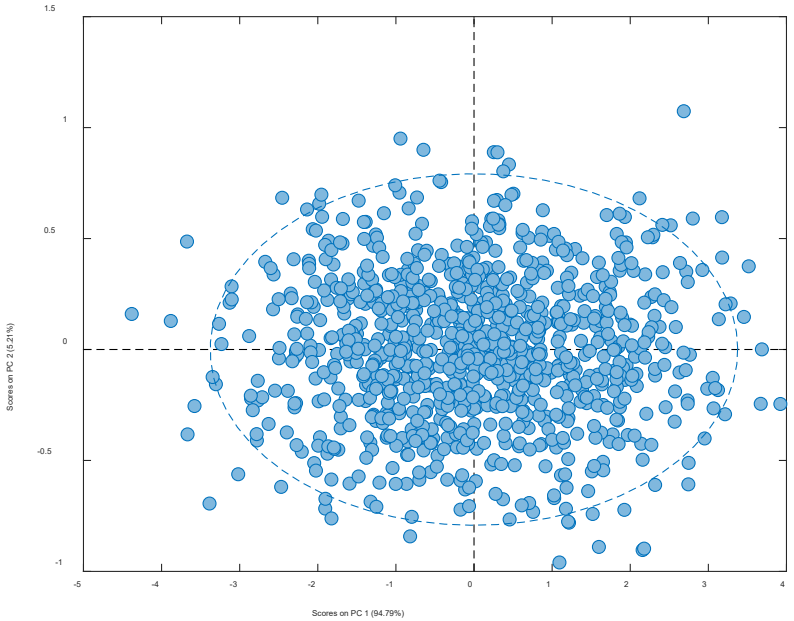
PCA monitoring model building

calibration and validation

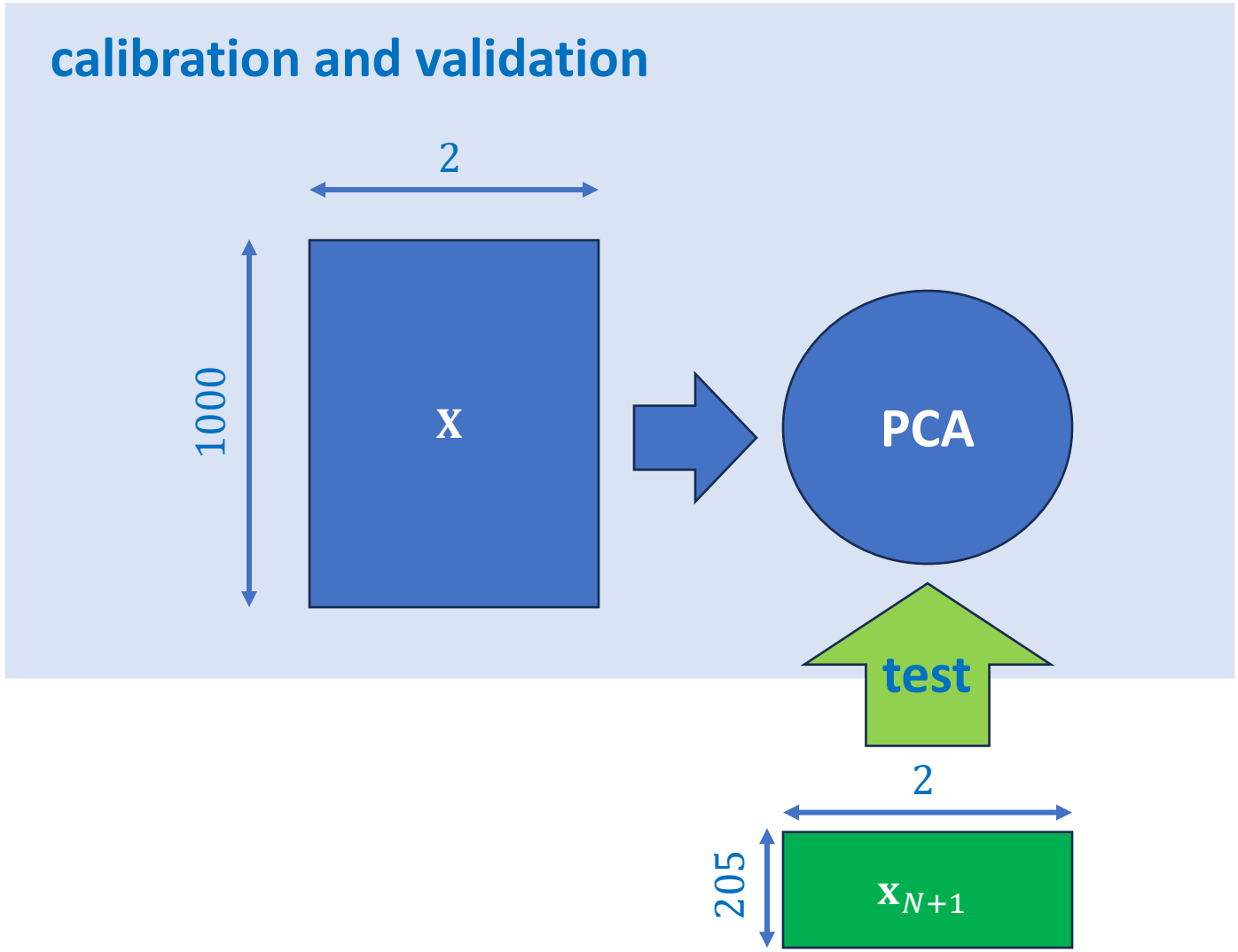


Simple process engineering case study

- A PCA model on 2 PCs can be built
  - PC1 explains almost all (~95%) of the variability, the systematic part
  - PC1 explains the positive and high correlation among the 2 variables
- The confidence limits in the score plot defines a monitoring chart which identifies the standard operating conditions:
  - 95% confidence limits leaves out 49 observation out of 1000, which is close to the expected 5%

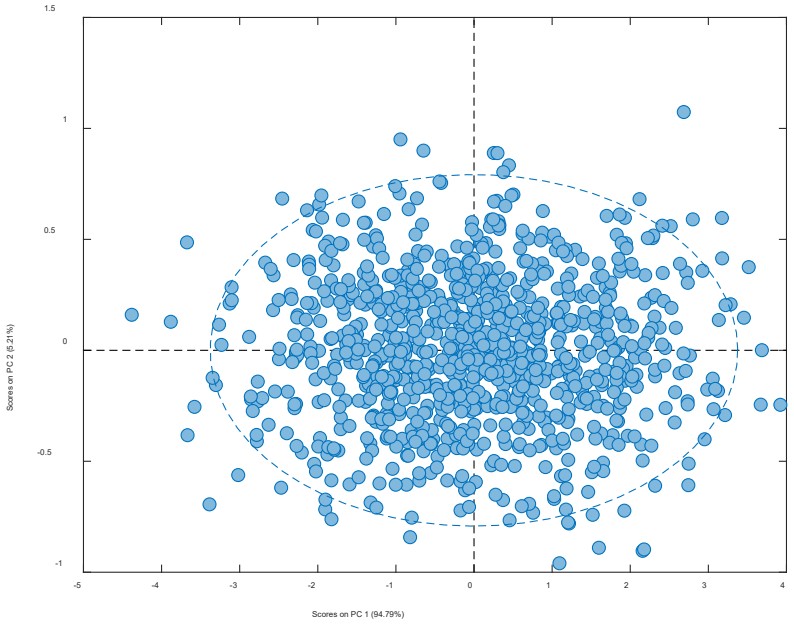


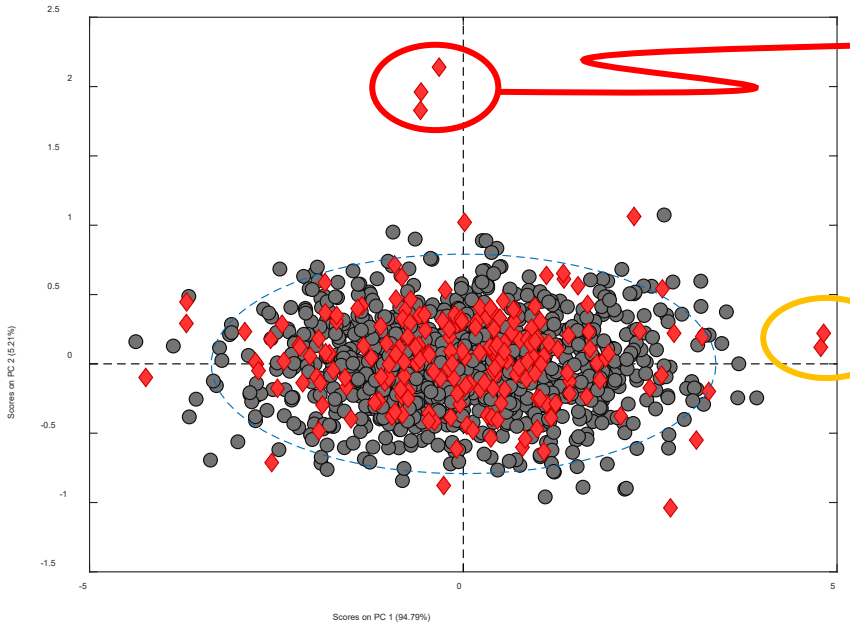
PCA monitoring model building



Simple process engineering case study

- A PCA model on 2 PCs can be built
  - PC1 explains almost all (~95%) of the variability, the systematic part
  - PC1 explains the positive and high correlation among the 2 variables
- The confidence limits in the score plot defines a monitoring chart which identifies the standard operating conditions:
  - 95% confidence limits leaves out 49 observation out of 1000, which is close to the expected 5%
  - 10 test observations (out of 200) are out even if they are normal operating conditions
    - 10 false alarms are warned
  - identify outlies



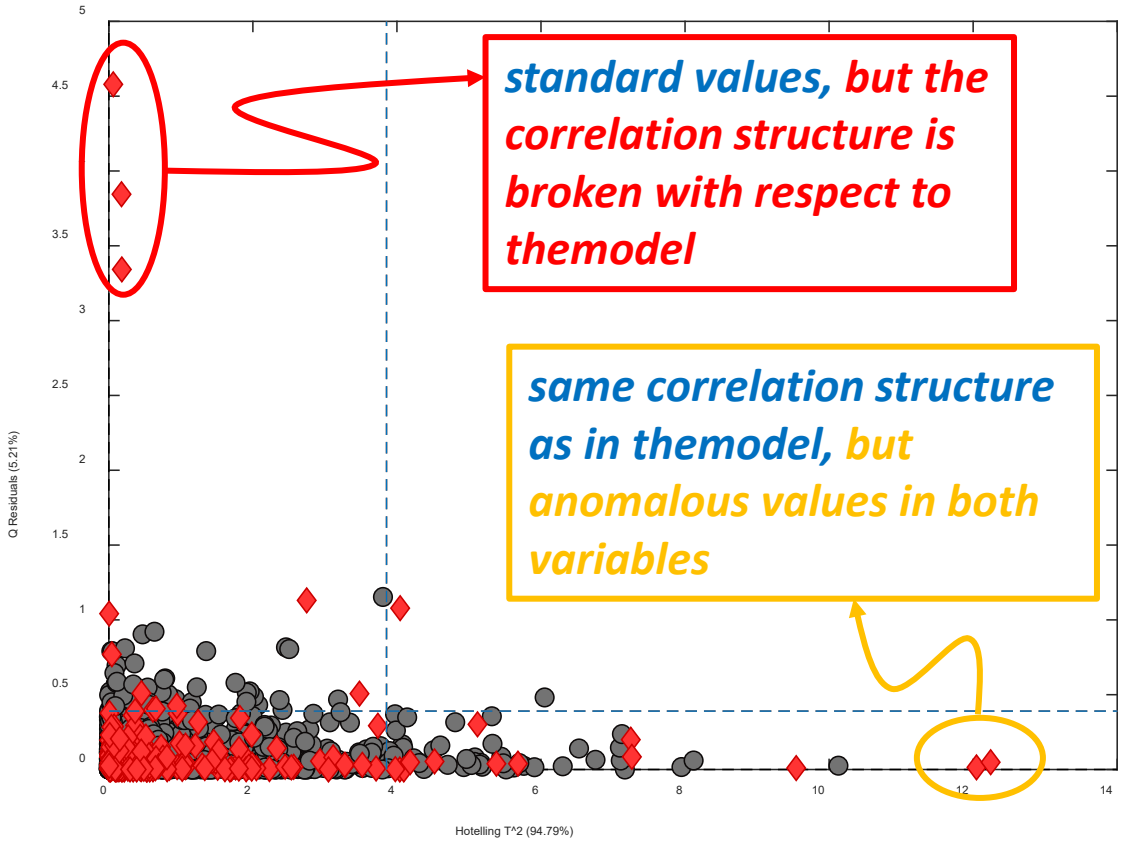


3 observations have standard values, but a different correlation with respect to the one described by PC1

2 observations share the same correlation structure as the one described by PC1, but show high values of both variables

Simple process engineering case study

- A PCA model on 1 PC captures the greatest part of variability:
  - the systematic part is retained
  - the positive correlation among variables is captured
  - noise is left out of the model space
- The 95% confidence limits
  - warn 5% for regular observations in both Hotelling T^2 and SPE
  - identify well as strong outliers the observations whose correlation structure



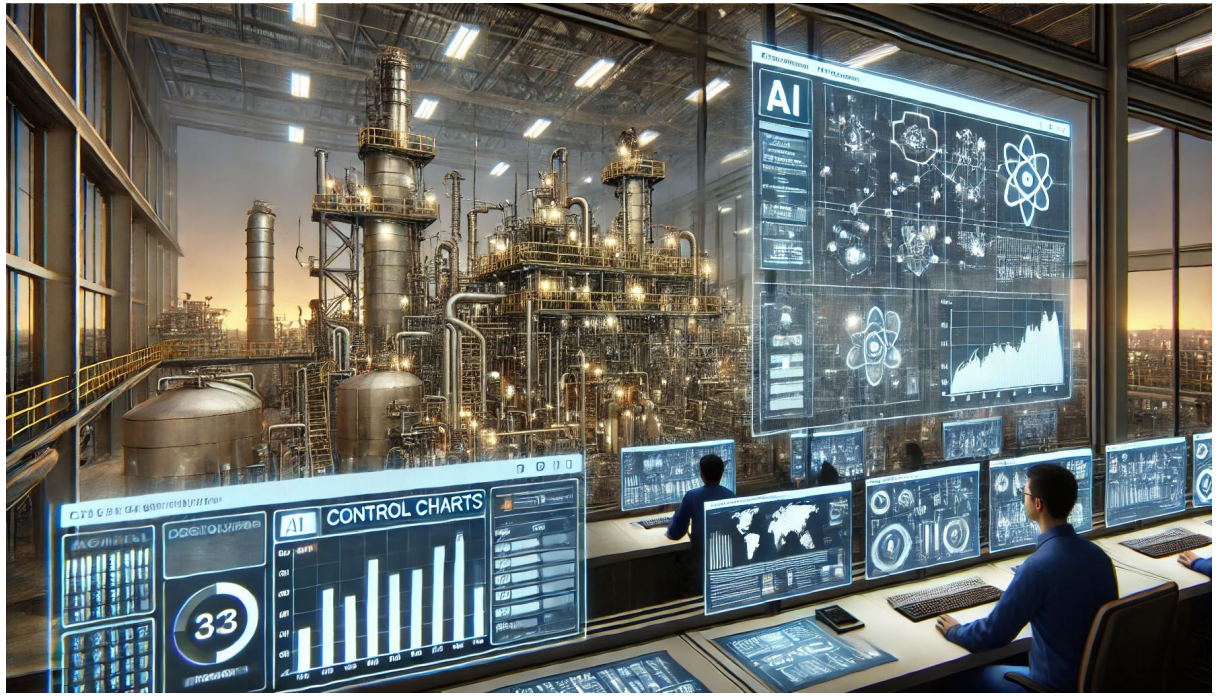
Today's homework

- Verify where these information are in the model structure generated by the PLS_Toolbox™
- Train with commands:
  - `tsqlim`
  - `residuallimit`
    - `jmlimit`
    - `chilimit`
  - `tconcalc`
  - `qconcalc`
- Code your own software to verify the formulation of the confidence limits and compare the results with the ones toolbox!
  - `chi2inv`
  - `finv`

Take-home message

- When a lot of correlated process or product quality data are available, multivariate process/quality monitoring systems can be built through latent-variables projection methods

- **Multivariate control charts** are used for monitoring purposes:
  - only two statistical indices, Hotelling T^2 and SPE, are used to surveille the process in cases where univariate charts fail



… per sempre a fianco a me!