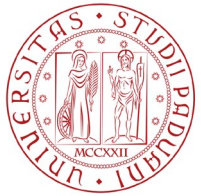


UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF INDUSTRIAL ENGINEERING

Machine Learning Lesson #11

Academic year 2025-2026

Prof. Pierantonio Facco

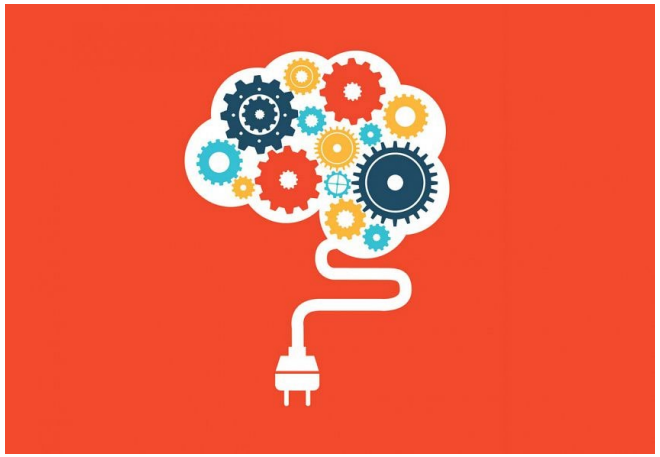
CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

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

Warm-up

- Please, connect to Kahoot!



Recap of the previous lectures

- PCA is an unsupervised learning methodology which can be used both for descriptive purposes and diagnostic purposes:
  - **descriptive analytics**: exploratory and correlative analysis
    - data dimensionality reduction
    - process understanding
    - process troubleshooting
  - **diagnostic analytics**: process and quality monitoring
- **Process and quality monitoring** is performed through multivariate monitoring charts (i.e., multivariate hypothesis testing)
  - Hotelling T^2 and SPE are utilized to detect anomalies, malfunctions and faults
  - fault causes' diagnosis is carried out through the contribution of the variables to T^2 and/or SPE

Today’s lesson

- Multivariate quality monitoring
- Multivariate process monitoring
- Examples:
  - copper production
  - continuous chemical process

Example of copper quality monitoring

Monitoring the quality of cuprum Cu

- AIM: **monitoring the quality** of final copper product in Boliden AB (Sweden)

- Datasets:
  - 730 observations
    - 720 calibration observations
    - 10 test observations
  - 9 variables
    - impurities in the produced Cu
    - purity quality index: TAI
      - TAI < 8: good product
      - TAI > 8: product to be discarded
      - **TAI = 8 is a specification limit**

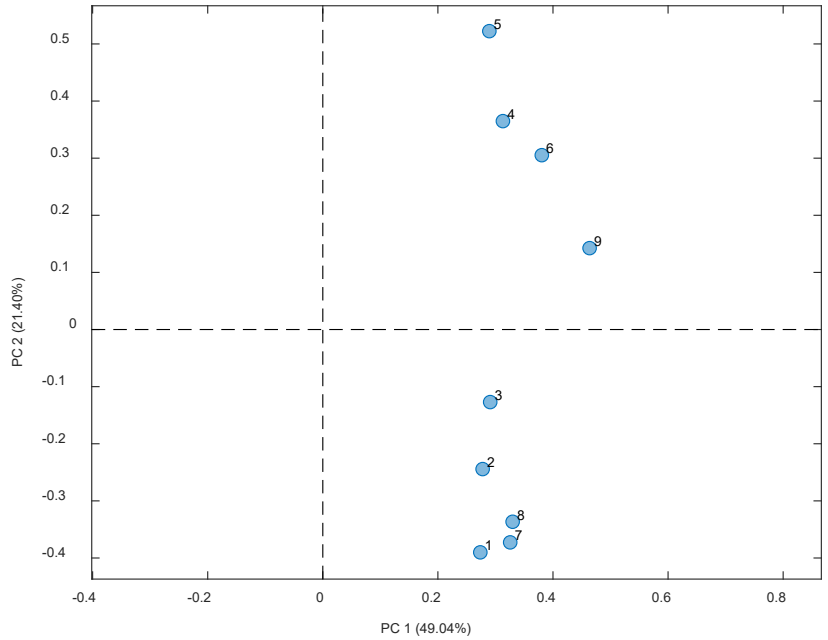
	variables
1	kAg
2	kNi
3	kPb
4	kBi
5	kSb
6	kAs
7	kTe
8	kSe
9	TAI

Copper quality PCA model

- After data pretreatment, a PCA model is built
- Loadings:
  - all the variables are positively correlated on PC1
    - the concentrations of all the contaminants increase together for ~50% of the observations
  - variables 4, 5, 6, and 9 are anti-correlated to the other ones in more than 20% of the observations

PC	eigenvalue	exp. variance (%)
1	4.41	49.04
2	1.93	21.4
3	0.76	8.29
4	0.68	7.58
5	0.48	5.35
6	0.41	4.57

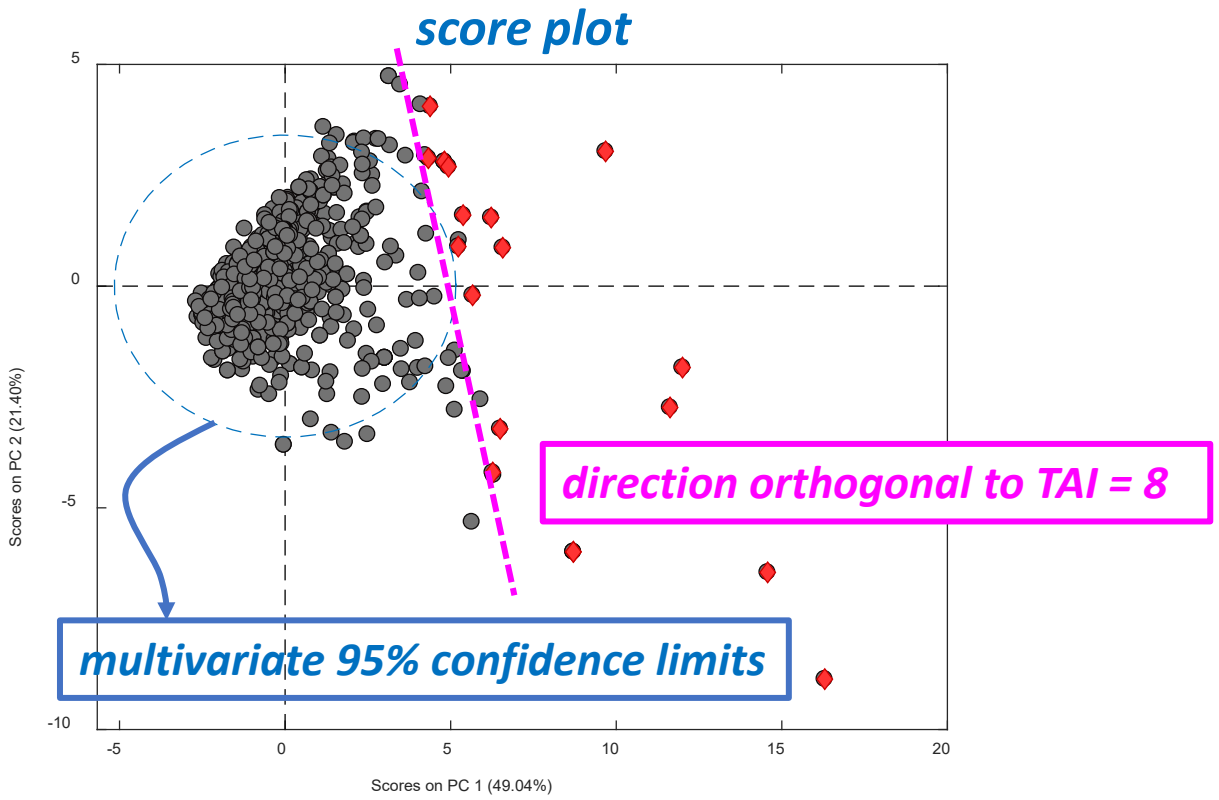
loading plot

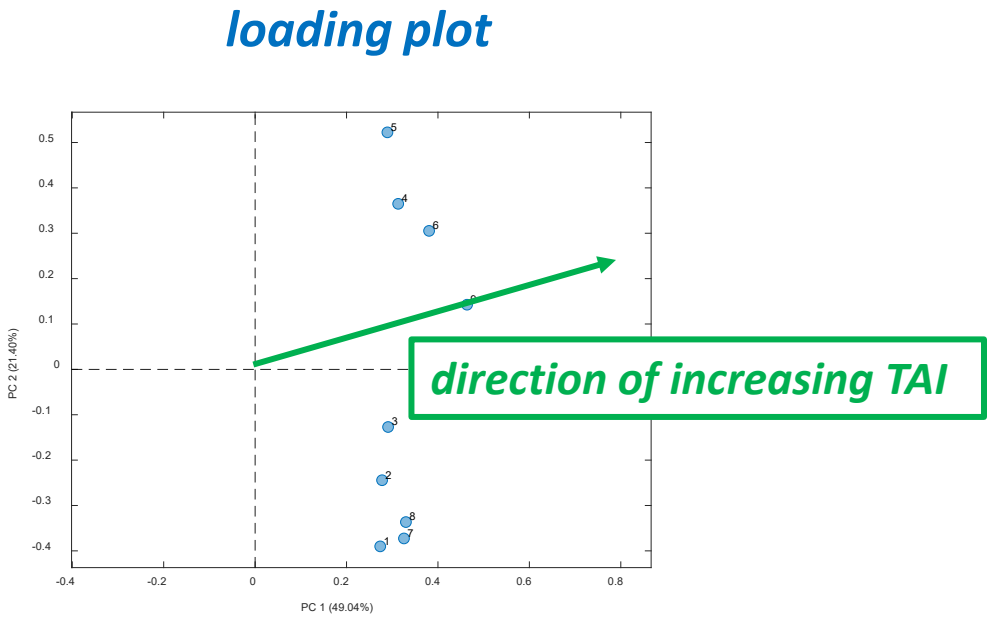


	variables
1	kAg
2	kNi
3	kPb
4	kBi
5	kSb
6	kAs
7	kTe
8	kSe
9	TAl

Monitoring bad copper quality

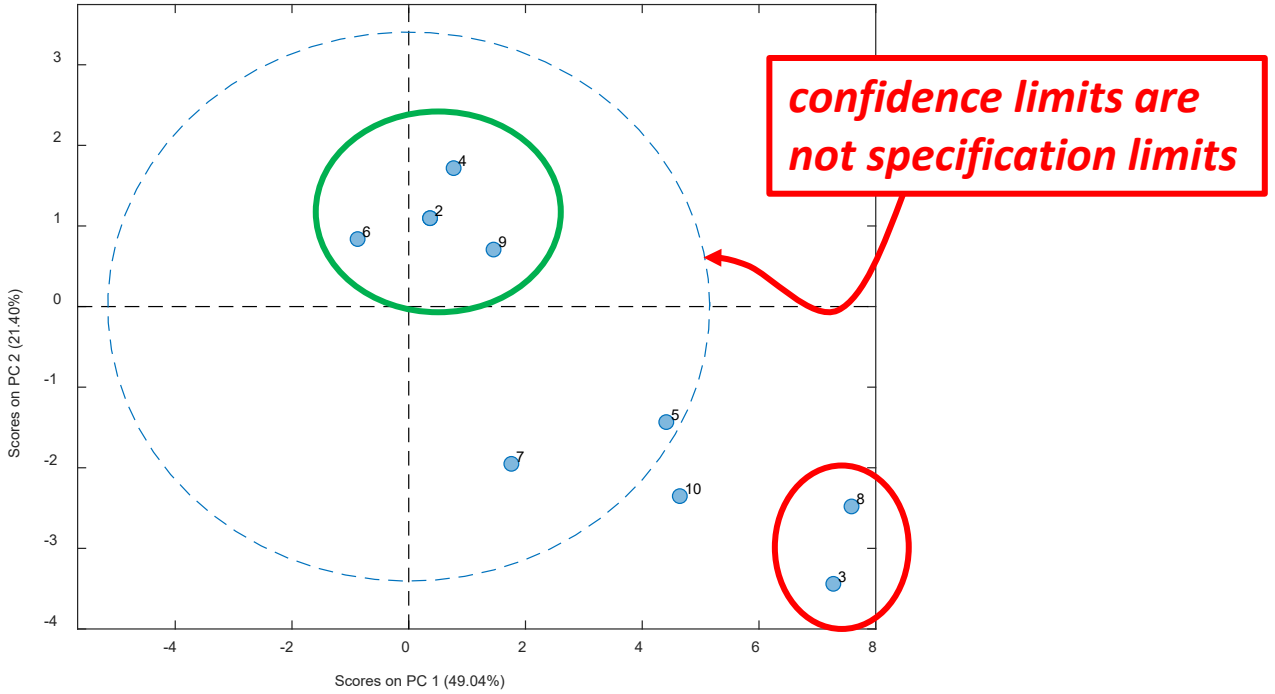
- Score monitoring chart (with 99% confidence ellipse):
  - bad products (TAI > 8) in the calibration dataset are highlighted in red
  - the contaminated products are projected into the score space of high PC1





New copper products

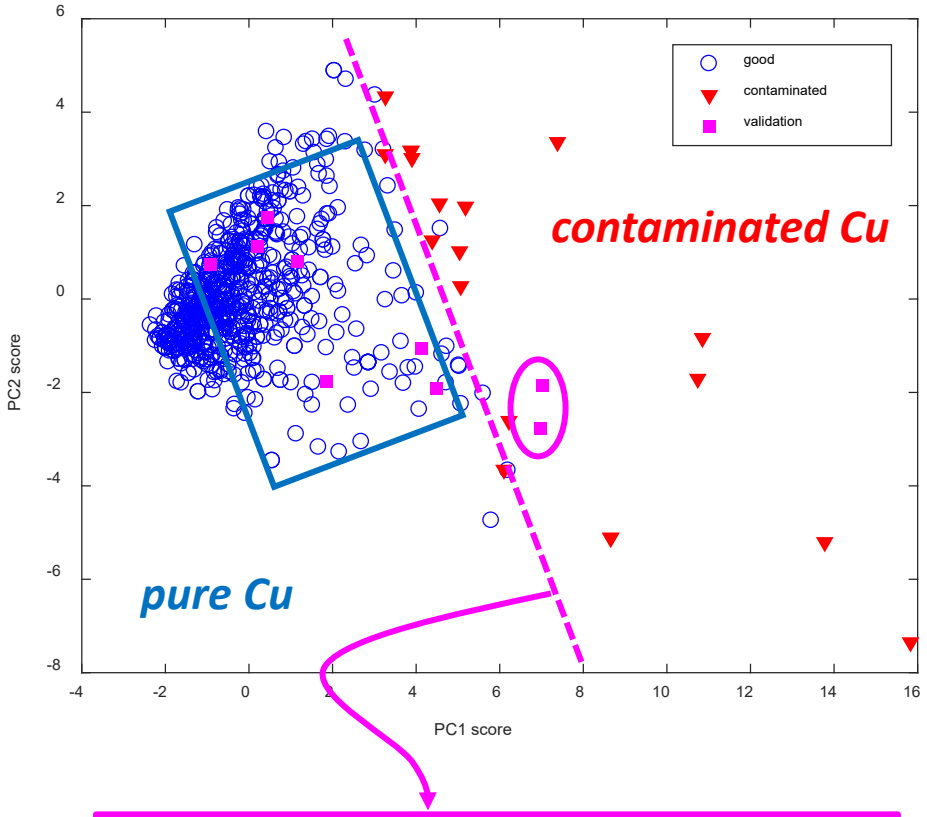
- 10 new observations (validation data, unknown to the PCA model) are projected onto the score space
  - observations **2, 4, 6, 9** are **good products** close to the average
  - observations 5 and 10 attribution to bad or good products seems to be uncertain
  - observations **3 and 8 seem to be anomalous**
    - in fact, TAI>8



validation observation	TAI
1	4.21
2	4.21
3	8.676667
4	4.736667
5	6.933333
6	3.373333
7	4.75
8	9.346667
9	5.153333
10	6.786667

Classification problem?

- The same problem could be seen also as a **clustering/classification/regression problem**:
  - PCA is an unsupervised methodology
    - can be used to observe data clustering
  - the **score plot of the PCA model built on all the classes** is a clustering map identifying the zones of the score space that pertain to each product class
    - good products used for model calibration (blue circles o)
    - bad products used for model calibration (red triangles ▼)
  - the joint reading of the score plot with the **loading plot** shows the variables that are most related to the **direction of clusters' discrimination**
  - the projection of new unknown samples (magenta squares ■) into the scores space may attribute the class to the new observations
  - the boundary among clusters does not have any "statistical" meaningfulness



univariate specification limits are projected into the score space as lines

Example of a continuous process monitoring

Continuous process monitoring

- Information on the continuous process under study is crypted due to confidentiality reasons

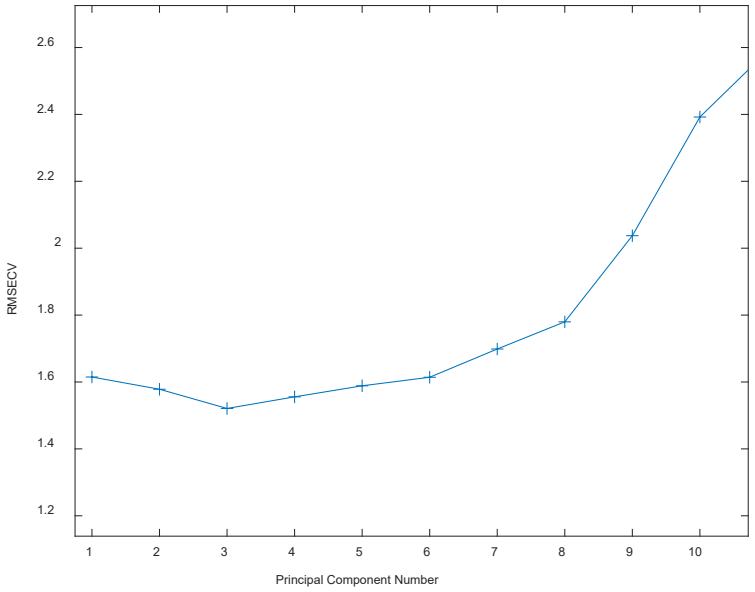
- AIM: **process monitoring**

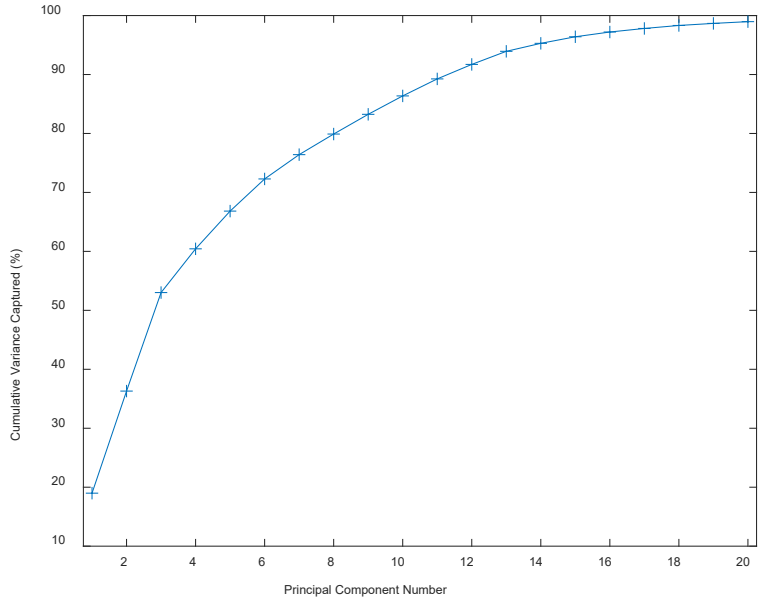
- Dataset:
  - observations:
    - 65 observations on NOC
    - 27 new observations
  - 29 online measured variables (not disclosed)
    - controlled variables
    - online measured variables
    - online measured output variables

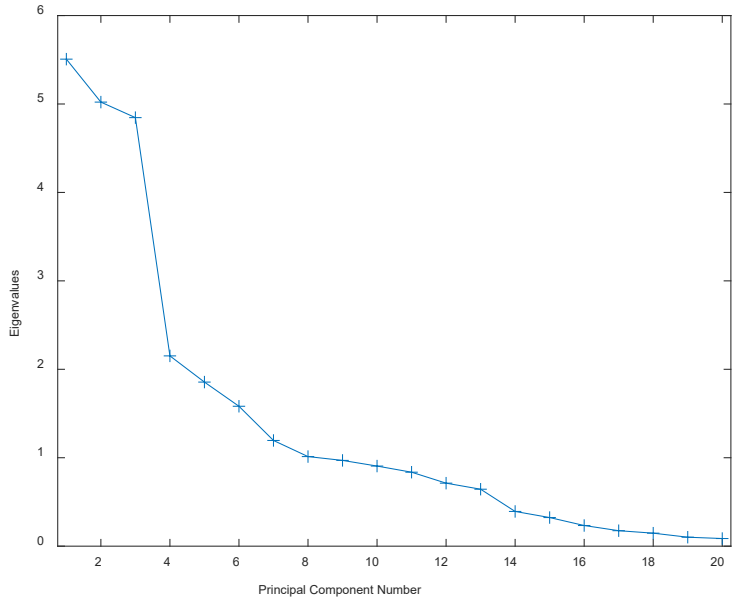
PCA monitoring model structure

- A PCA model was built:
  - 3 PCs are selected explaining 53% of the data variability

Are we worried that this is low?







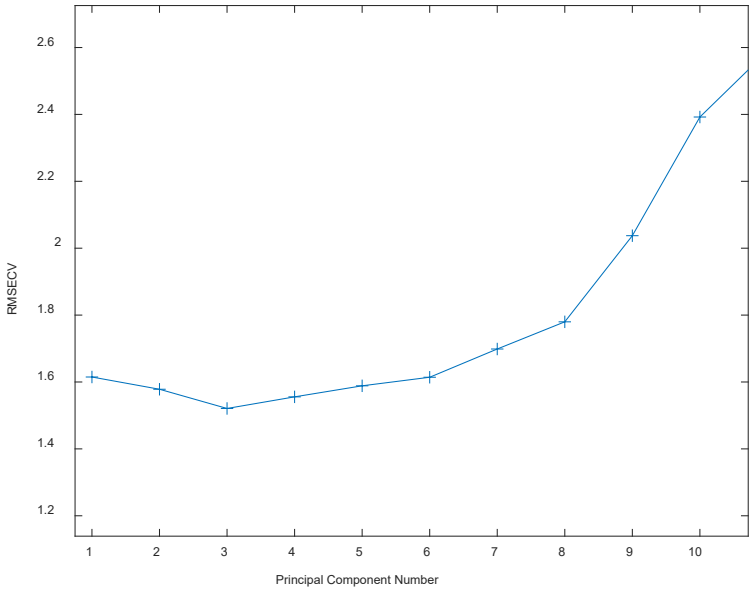
Group discussion

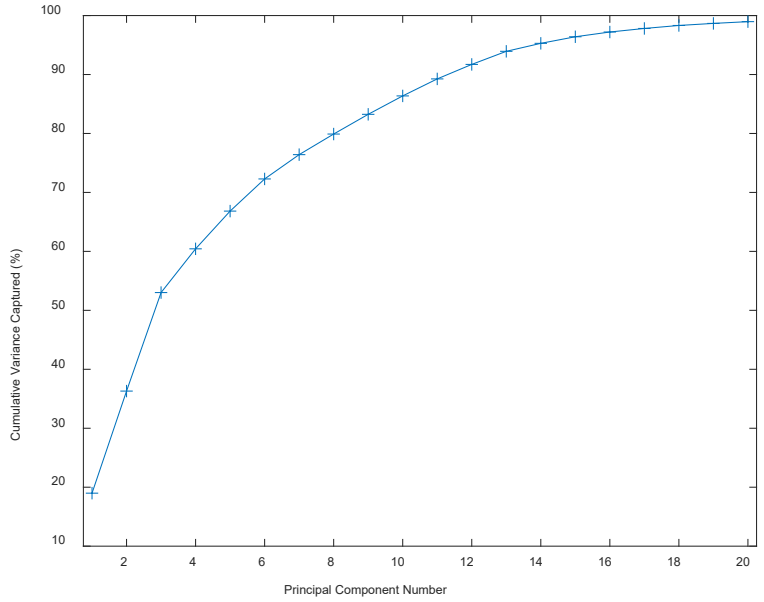
- Are we worried if a monitoring model explains a low percentage of the available data variability?

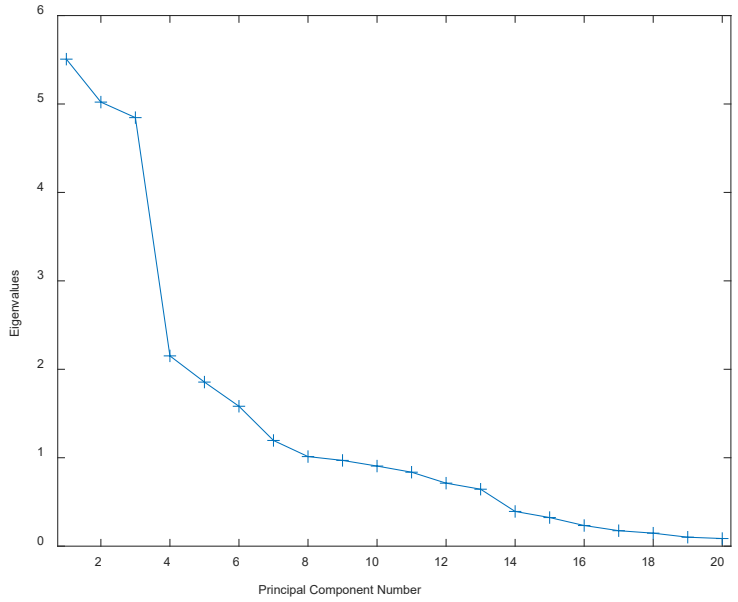
PCA monitoring model structure

- A PCA model was built:
  - 3 PCs are selected explaining 53% of the data variability

Are we worried that it is low? No, not in process monitoring! The remaining 47% is represented by the squared prediction error whose chart is considered for monitoring

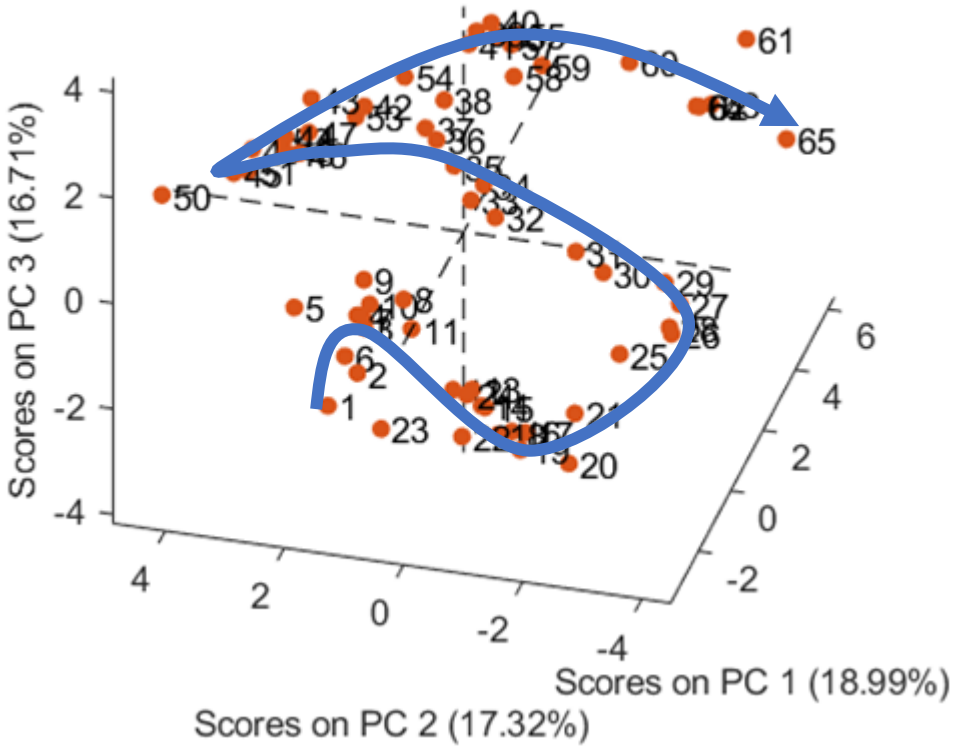






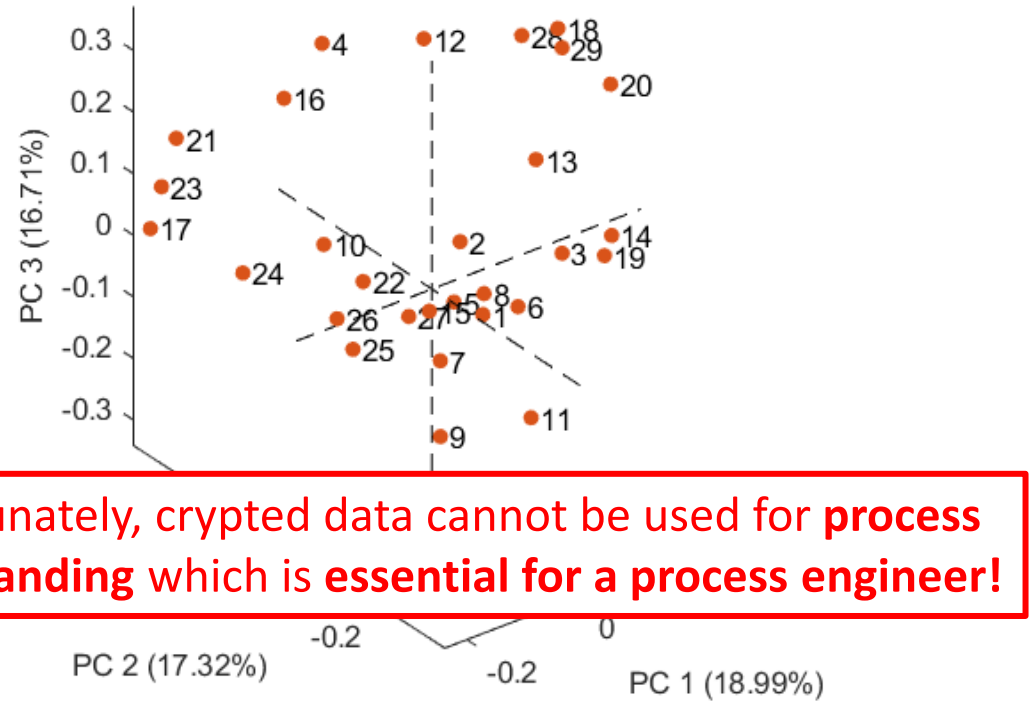
Relation among observations

- A lot of **auto-correlation** is present in the data
  - one observation is related to the previous and the following ones
  - the process seems to have a sort of “memory” of what happened in the previous observations

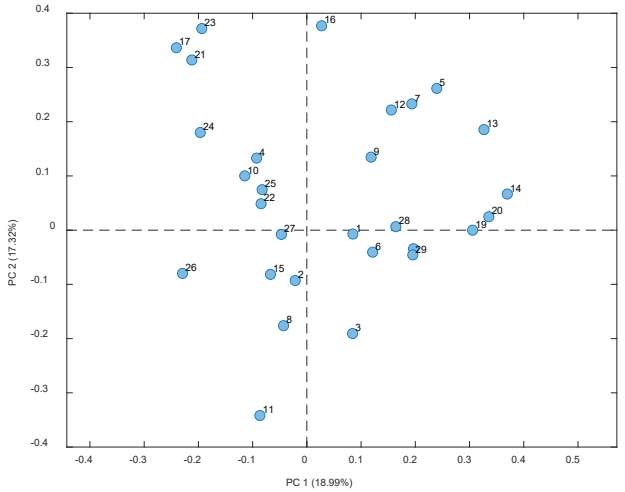


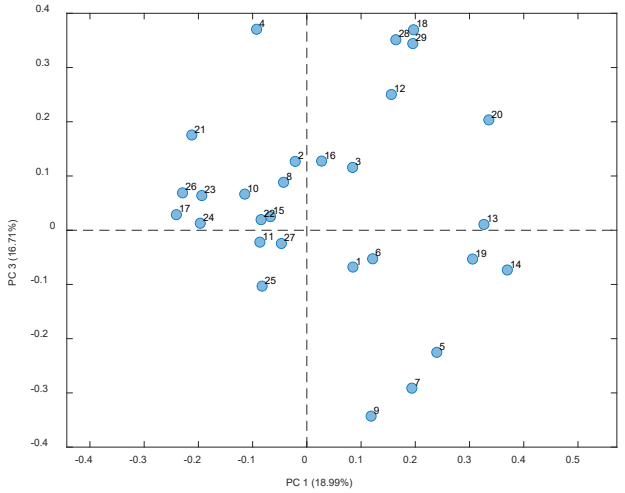
Correlation between variables

- The loading plot shows that:
  - the first 3 PCs explain almost the same aliquot of variability
  - the most influential variables are:
    - 13, 14, 19, 20, anti-correlated to 17, 21, 23, 24, 26 on PC1
    - 16, 17 21, 23, anti-correlated to 11, etc. on PC2



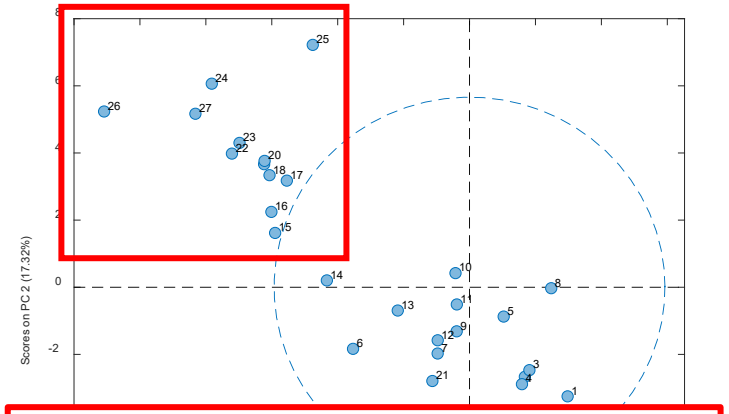
unfortunately, crypted data cannot be used for **process understanding** which is **essential for a process engineer!**



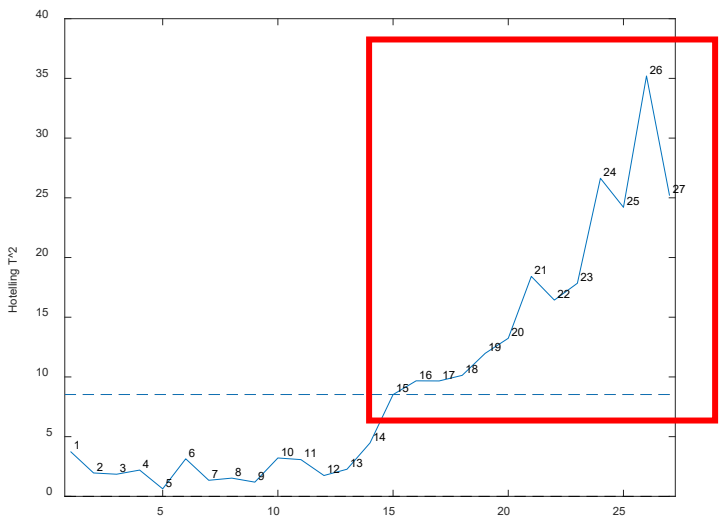


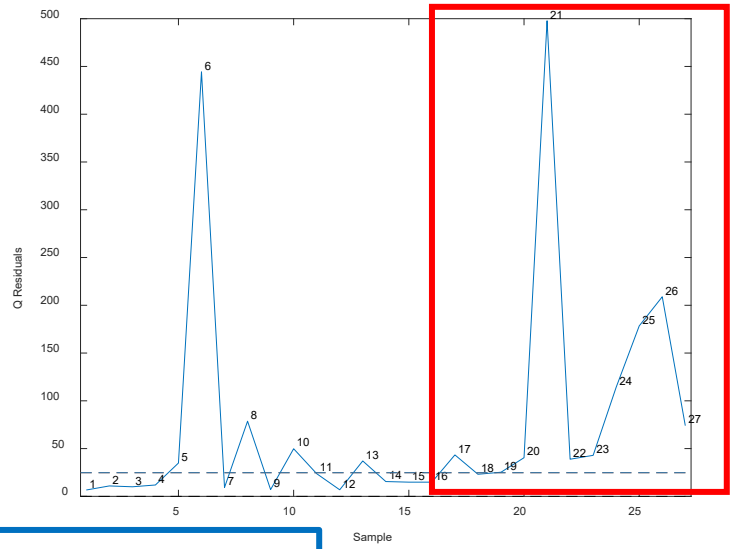
Anomalies detection for new observations

- New observations are projected onto the PCA model
  - the Hotelling T^2 and the squared prediction error Q are observed
  - from observation 15 to 27 a process drift is detected
  - the difference from the average conditions increases with time
  - the correlation structure is completely broken after few (i.e., 6) observations



the PC1 vs. PC2 score plot does not tell the entire story because the model is composed of 3 LVs

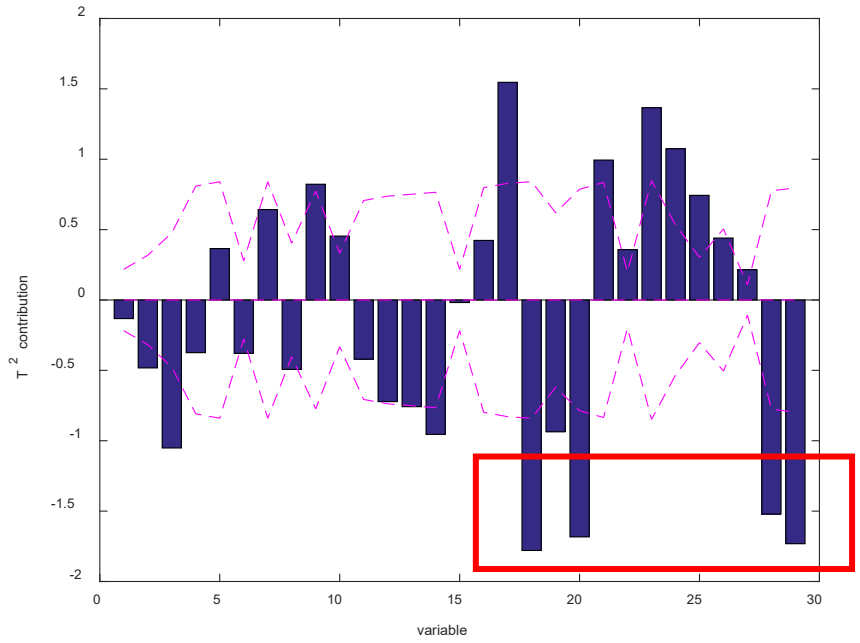


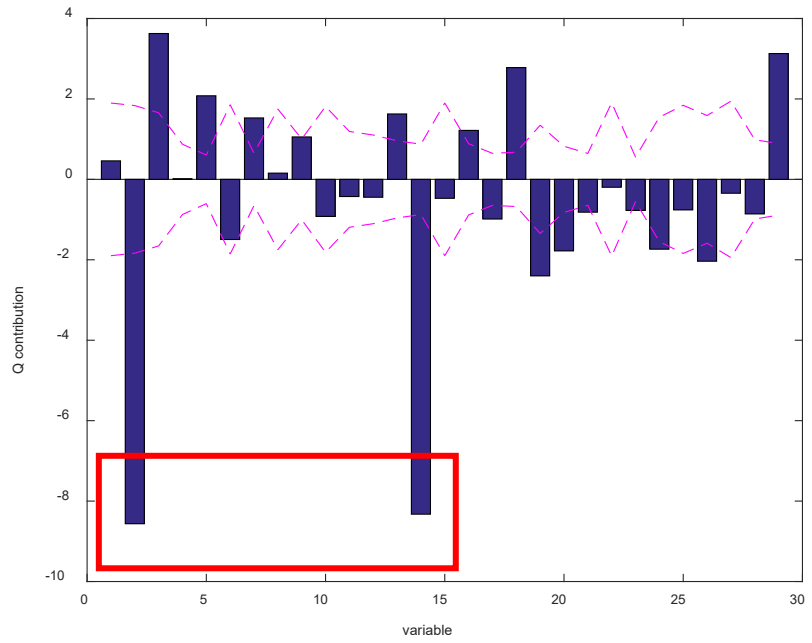


the T^2 and Q statistics does tell the entire story!

Diagnosis of the problems

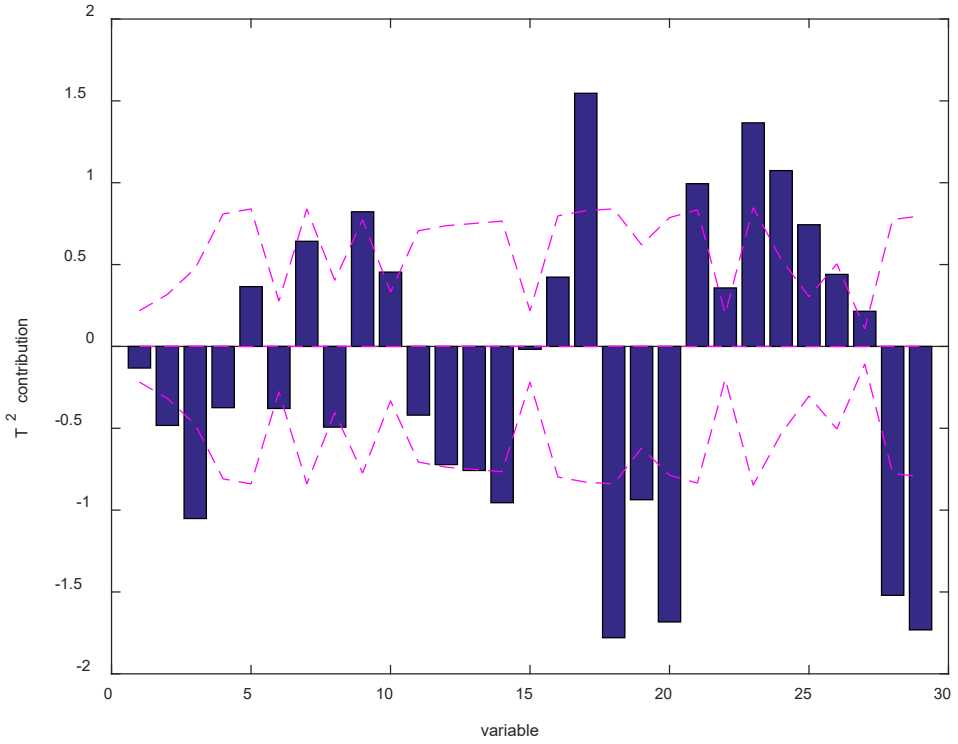
- The diagnosis is completed by comparing the **contributions to the Hotelling statistics** and the **contributions to the square prediction error** to the respective “standard” contributions in the calibration dataset
- Here the contributions to T^2 and Q are shown for observation 26
  - the variable that seem to be mostly related to the malfunction are:
    - 18, 20, 28 and 29 (too low with respect to the standard in calibration) and 17 and 23-25 (too high) in the T^2 contributions
    - 2 and 14 broke the correlation structure with the other ones in the Q contributions





Contribution plots reading

- Both the **contributions to the Hotelling statistics** and the **contributions to the square prediction error** are distributed in a Gaussian fashion
- A 95% confidence limit can be calculated for each contribution to both T^2 and Q from a Gaussian distribution from the calibration data



Procedure for PCA process monitoring

select the proper number A of PCs

- consider eigenvalues >1 (or those >0.7/0.8)
- identify the knee on the cumulative explained variance or the elbow in the explained variance
- trust the minimum of the RMSE in cross-validation (appropriately considering the scale of the error)
- CROSS THE ABOVEMENTIONED INFORMATION AND DO NOT FORGET TO COMMENT THE **AMOUNT OF EXPLAINED VARIANCE**

comment scores

- comment multinormal behavior
- identify clusters
- identify patterns, such as auto-correlation
- put an ellipse only if the observations are multinormally distributed
- study if the rate of outliers from confidence ellipse is the expected one
- remove outliers only if there is a special cause and rebuild the model

comment loadings

- identify correlation
- identify anti-correlation
- identify independence
- DO NOT FORGET TO COMMENT THE IMPORTANCE OF THE EXPLAINED VARIABILITY AND **TRANSLATE THE OUTCOME AS A PROCESS EXPERT**!

monitoring charts

- consider both the hoteling statistics and the Q residuals
- evaluate if the number of outliers is the expected one (5% for 95% confidence limits)
- project test observations
- warn an alarm if one observation is out of the limit of at least one chart
- if an alarm is warned diagnose the cause through the relevant contribution plot with the respective limits

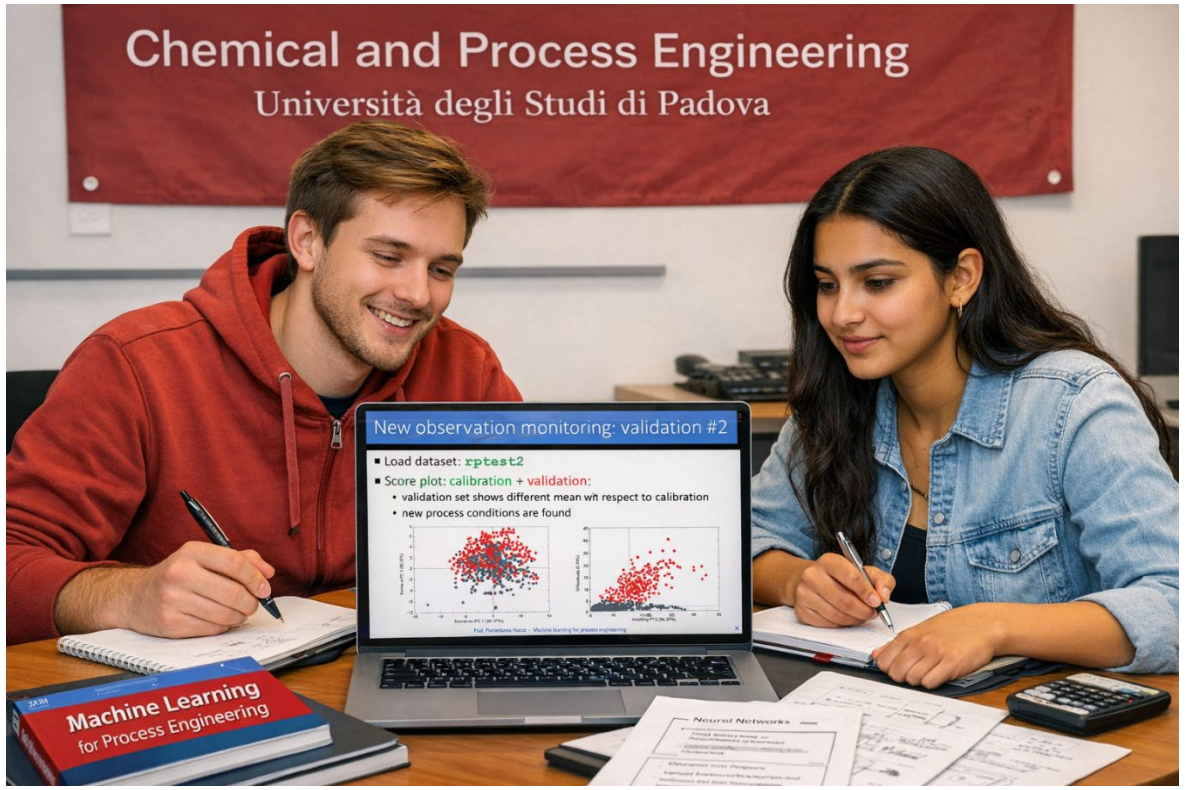
Take-home message

- PCA is an unsupervised methodology that is commonly used for process and quality monitoring:
  - 2 **monitoring charts** are used to summarize all the information in a (large number of) original variables
    - chart on Hotelling T^2
    - chart on Q residuals
  - the monitoring charts are effective in the detection of departures from standard behavior of a plant/product both in terms of:
    - variability
    - correlation
  - **faults, malfunctions and anomalies** can be detected in a timely manner
    - the causes of faults, malfunctions and anomalies can be **diagnosed** through contribution plots

Today's homework

- Practice with Matlab® and PLS_Toolbox®
  - datasets are available in Moodle for both the examples of:
    - copper quality
    - continuous process

- See the video on the computational **Laboratory #2** in Moodle
  - you will learn how to build monitoring models
  - you will practice with a new dataset in the case of:
    - monitoring of a slurry-fed ceramic melter for the stabilization of nuclear waste



… per sempre a fianco a me!