Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #12

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 from the previous lessons

- **Multivariate statistical techniques** are key methodologies to consider the case of **multivariate correlated datasets**:
  - they project the data in a reduced set of latent variables
  - latent variables
    - describe the largest part of the data variability in the original data
    - provide an optimal fitting of the multivariate data
- **Data-based monitoring** methodologies are commonly based on PCA:
  - it uses multivariate statistical monitoring charts to **detect anomalies**, namely, to understand if data:
    - are similar to the data observed in a set of calibration samples from system/process NOC (**Hotelling control chart**)
    - conform to the correlation structure of those calibration NOC conditions (**squared prediction error control charts**)
  - it provides powerful **diagnosis** tools of the causes that determine the anomalies detected through the control charts (**contribution plots**)

Today's lesson

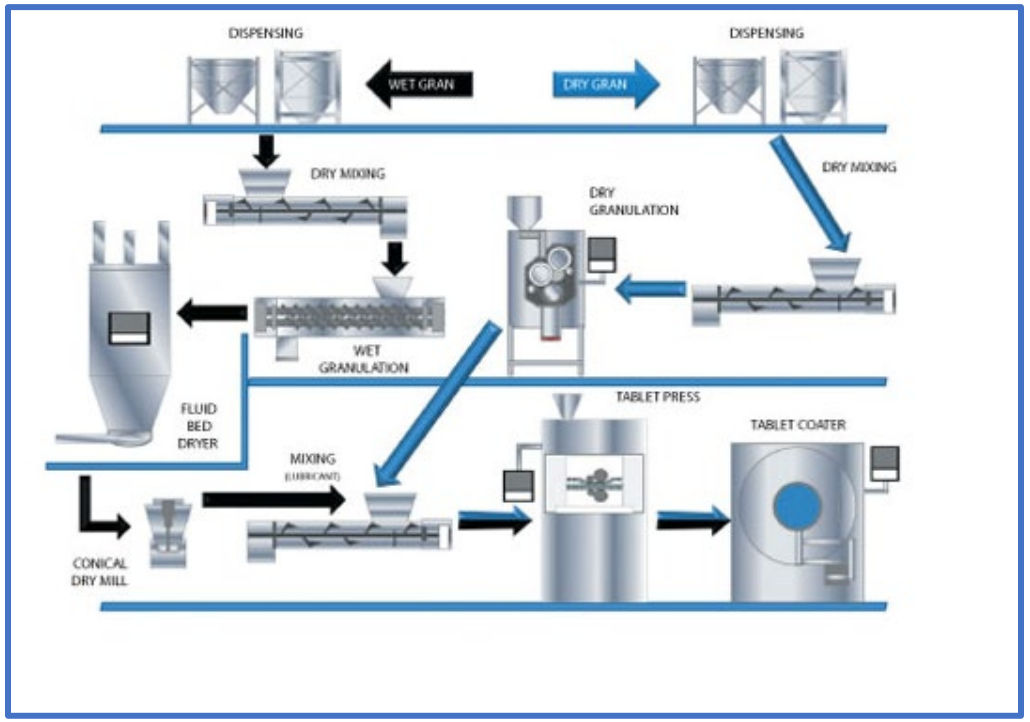
- Multivariate batch process monitoring
  - post-mortem monitoring
  - real-time monitoring

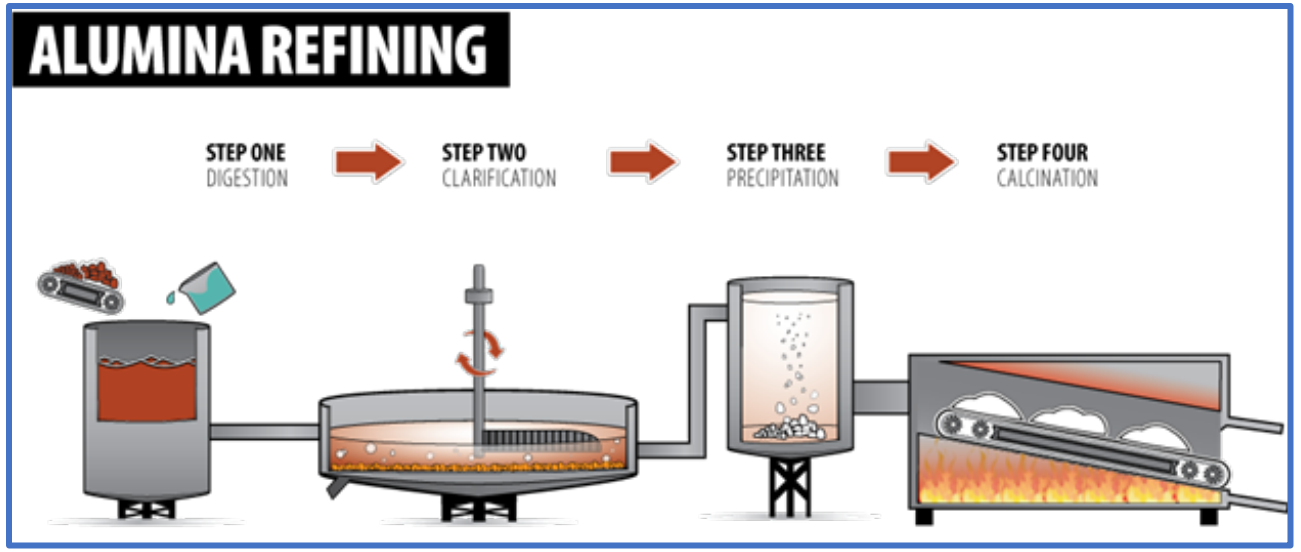
Continuous processes

- Until now we dealt with **continuous processes**, which:
  - have a constant flow of raw materials into production, generating a constant flow of products
    - is a non-stop production method
  - move raw material from the start of the process through each production step to a final product
    - raw materials are fed and processed continuously to manufacture additional units of product
  - **benefits**:
    - high production volumes
      - reduced processing and holding time
    - small storage space
    - easy process control
  - **challenges**
    - less flexibility
    - long to set up
    - risk in production startups and shutdown
    - high initial investment cost
    - high risk of contamination with products moving through the same process

Examples of continuous processes

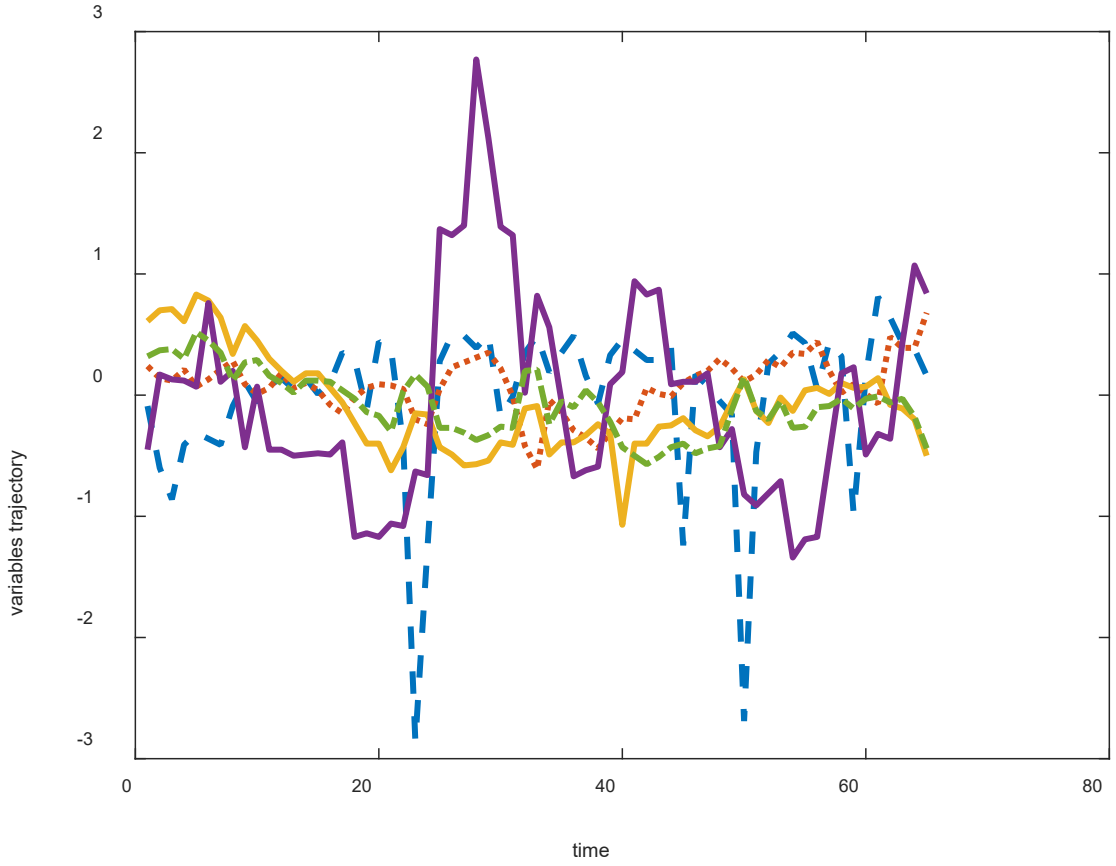
- Examples:
  - production of gasoline from oil in petrochemical industry
  - tablets production in the pharmaceutical industry
  - alumina refining process in mining industry, etc.





Continuous processes variables' time trajectories

- Variables time trajectories from a continuous process (autoscaled values)
  - only random variability is displayed
  - no time trajectories can be found
    - no trends in time
    - no trace of the operating stages



Batch processes

- A large part of the industrial processes (especially in chemical engineering) are **batch processes**
  - the production is carried out in different lots
  - the process follows a sequence of steps carried out in different operating phases:
    - initialization
    - raw materials load
    - …
    - product completion
    - product discharge
  - they have finite duration:
    - they are not continuous production processes
- Wide range of industrial applications:
  - pharmaceutical and bio-pharmaceutical industry
  - biotechnology sector
  - semiconductors manufacturing
  - plastic materials production
  - etc.

Benefits of the batch processes

- Small volumes of high value-added products are typically manufactured
- High production flexibility
  - different grades of products can be manufactured
- High control over quality
  - easier product traceability
- Relatively short production time
- Low equipment cost
- Low chance of contamination
  - all products move along the production process at the same time
- Batch processing is a well-established production method that is easier to manage and trace than continuous in highly regulated environments

Challenges of the batch processes (1/2)

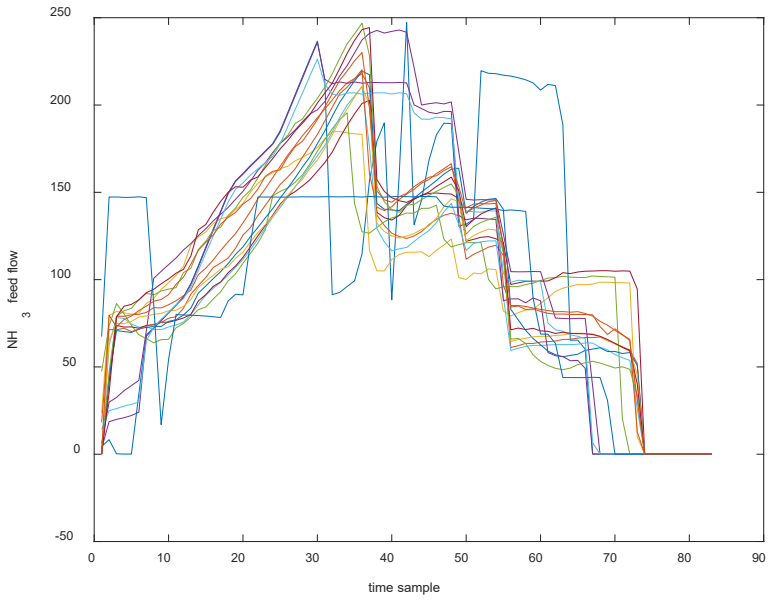
- From the **processing point of view**:
  - greater storage space needed
  - greater production rates of waste
  - higher production costs
  - increased employee downtime:
    - management of the process phases
    - meticulous quality control
  - unwanted disturbances may affect the entire batch or even a series of consecutive batches:
    - discards of low-quality products
    - reworks of a low product quality profile product
    - contamination of different products, waste, etc.

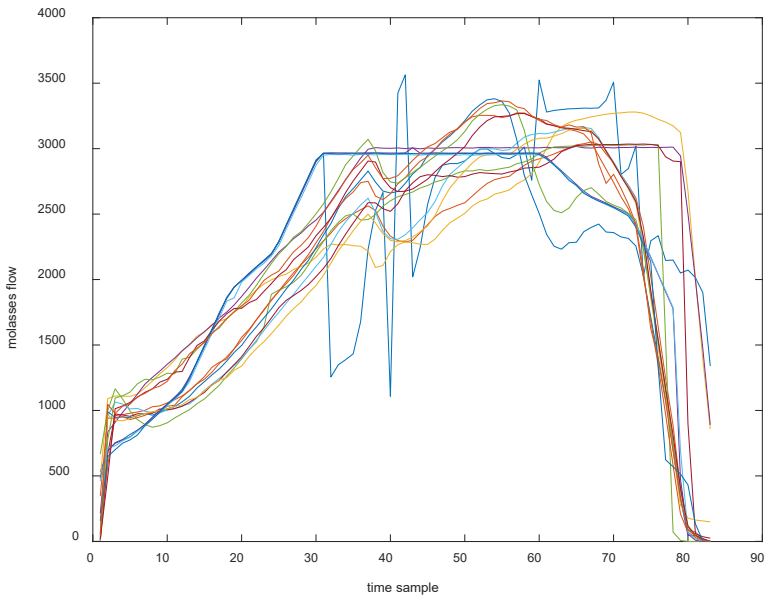
Challenges of the batch processes (2/2)

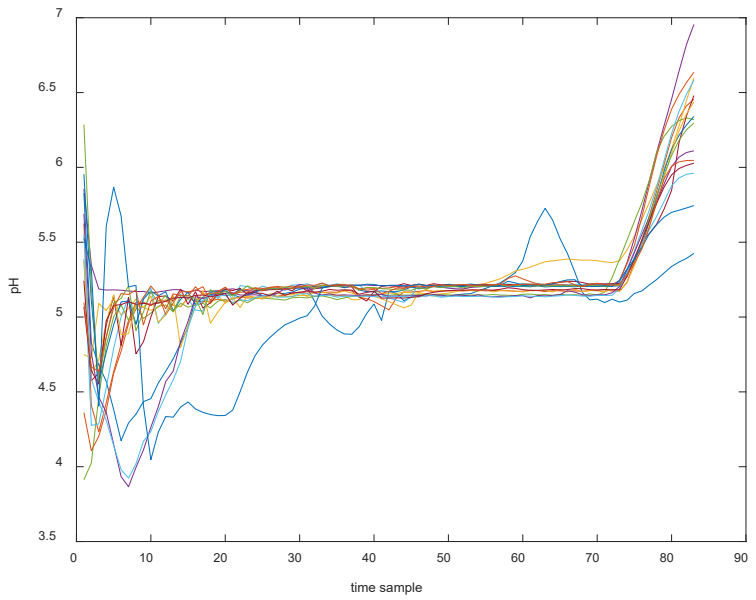
- From the **data point of view**:
  - variables display a **dynamic time dependency** that should be treated in an appropriate fashion
    - autocorrelation is present, and not only cross-correlation
  - large **variability of the operating conditions** and large **batch-to-batch variability**
    - deviation of the variables time profiles from the target ones
    - variability on the input materials
      - properties
      - impurities
      - different providers
    - differences in settings
    - manual operations from the operators
    - scarce automation

Batch processes variables' time trajectories

- Variables time trajectories from a batch process to produce baker yeast
  - trends in time
  - trace of different operating stages (among which process start-up and shut-down)







- Are you ready? Are you tough enough?

- Are you ready? Are you tough enough?

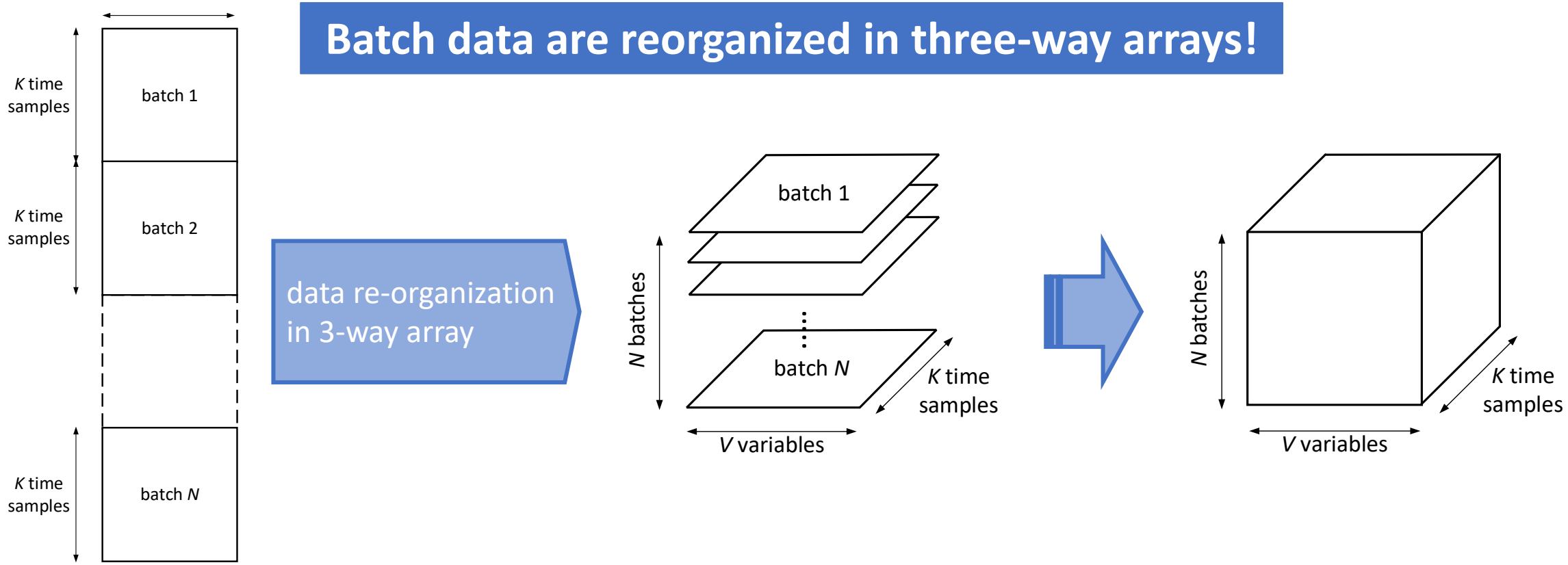
Multi-way data structure

Multi-dimensional data arrays and dealing with variability and auto-correlation in time

Multi-way data arrays

- A different **data structure** is present in batch processes:
  - typically, data are organized in tall arrays which contain the recordings of the variables time trajectories for different batches

Batch data are reorganized in three-way arrays!

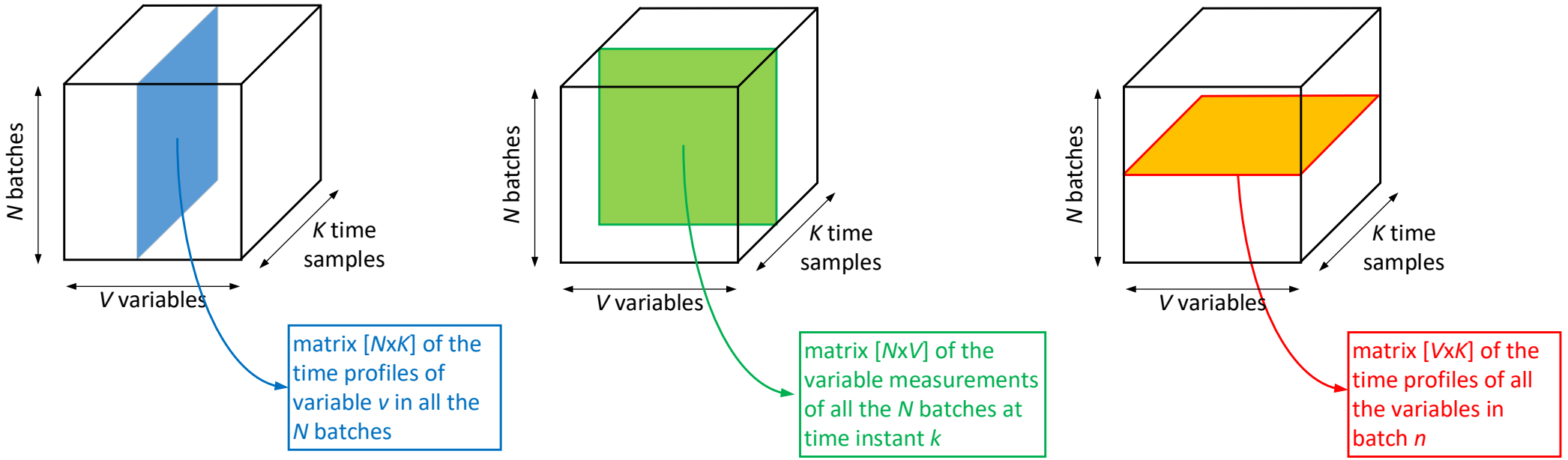


Structure of the 3-way batch data arrays

- The three-dimensional array is organized in a matrix:

$$\underline{\mathbf{X}} = [N \text{ batches} \times V \text{ variables} \times K \text{ time instants}]$$

- The three-dimensional matrix can be observed by **"slicing" the 3D $\underline{\mathbf{X}}$ matrix in different directions** (i.e.: ways)
  - direction of the variables:
    - all the time profiles of the same variable are analyzed for all the batches
  - direction of the batches:
    - all the variables at the same time instant are analyzed for all the batches
  - direction of the time:
    - all the time profiles of all the variables are analyzed for one batch



Three-dimensional data unfolding

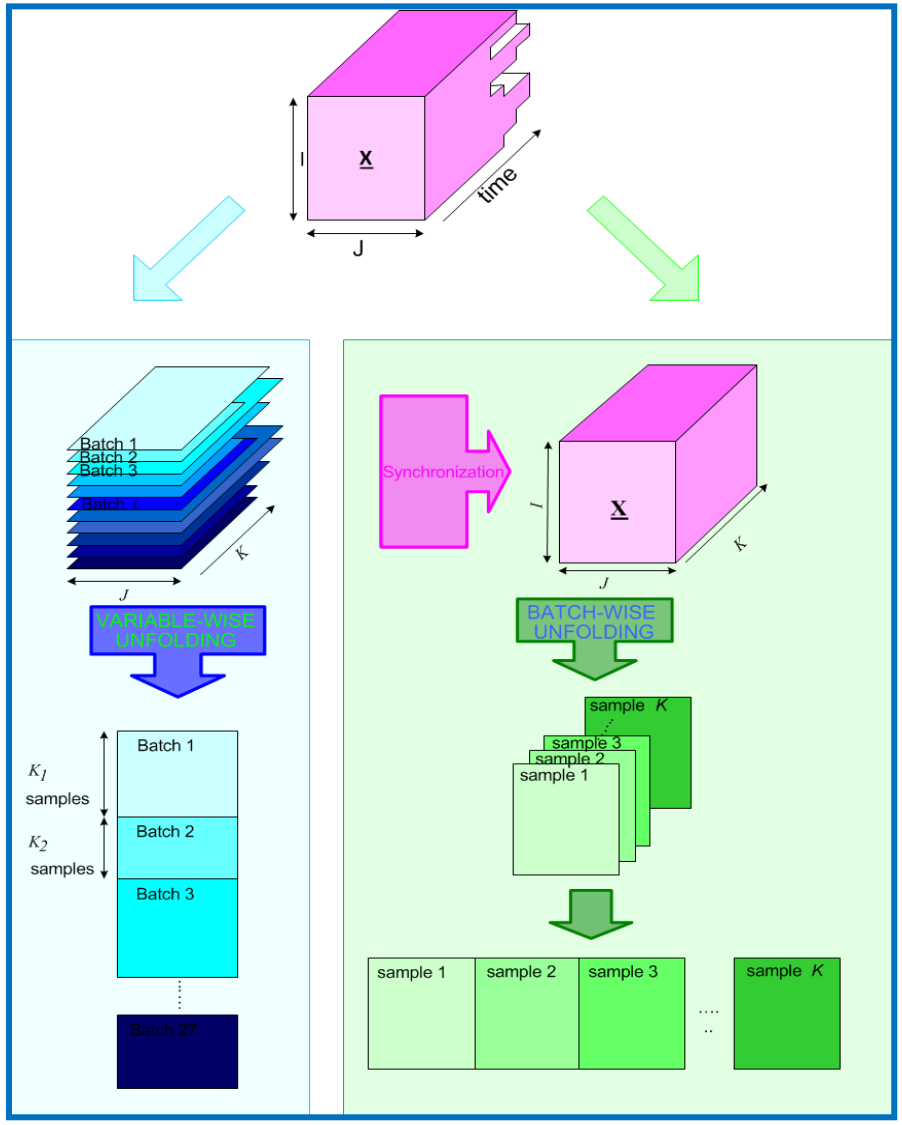
- Multivariate statistical techniques deal with bi-dimensional data
  - to treat 3D arrays the matrix must be reorganized in a bi-dimensional fashion

DATA UNFOLDING!

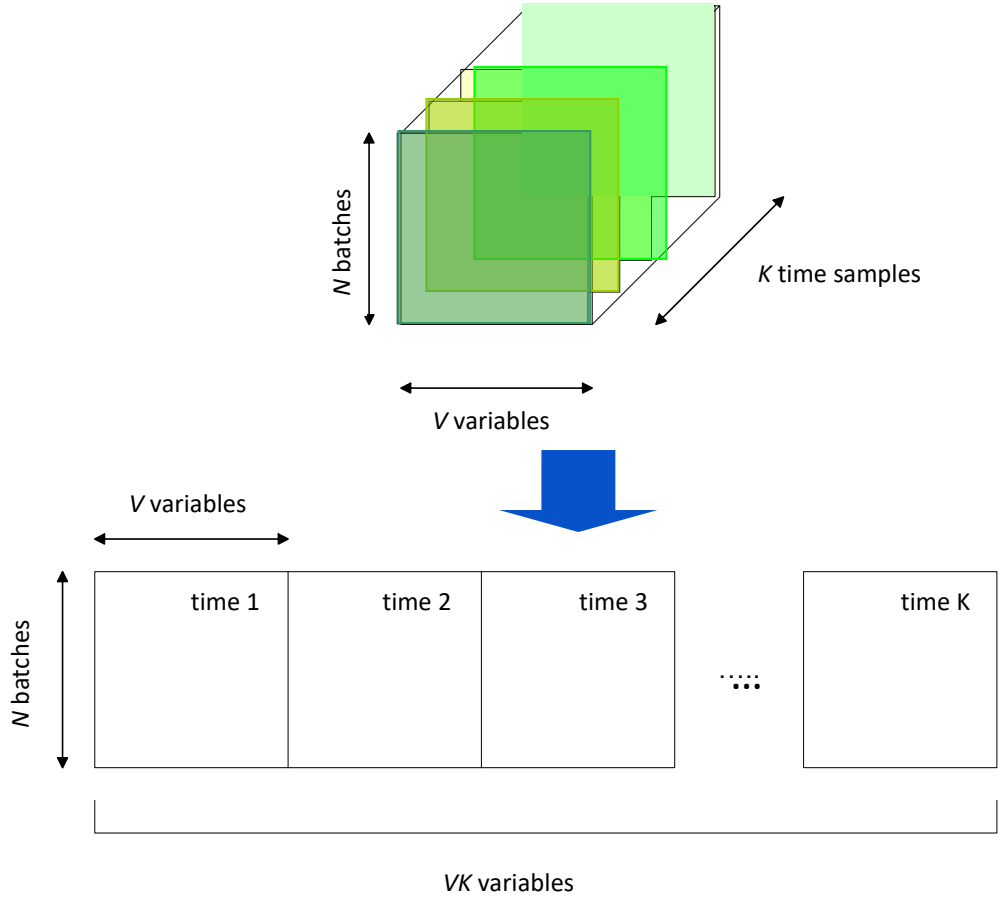
Discussion

- How could you propose to unfold the 3D data matrix $\underline{\mathbf{X}}$?
- How could $\underline{\mathbf{X}}$ be transformed in a 2D data matrix $\mathbf{X}$ that can be treated by multivariate statistical methods?

Unfolding methodologies



Batch-wise unfolding

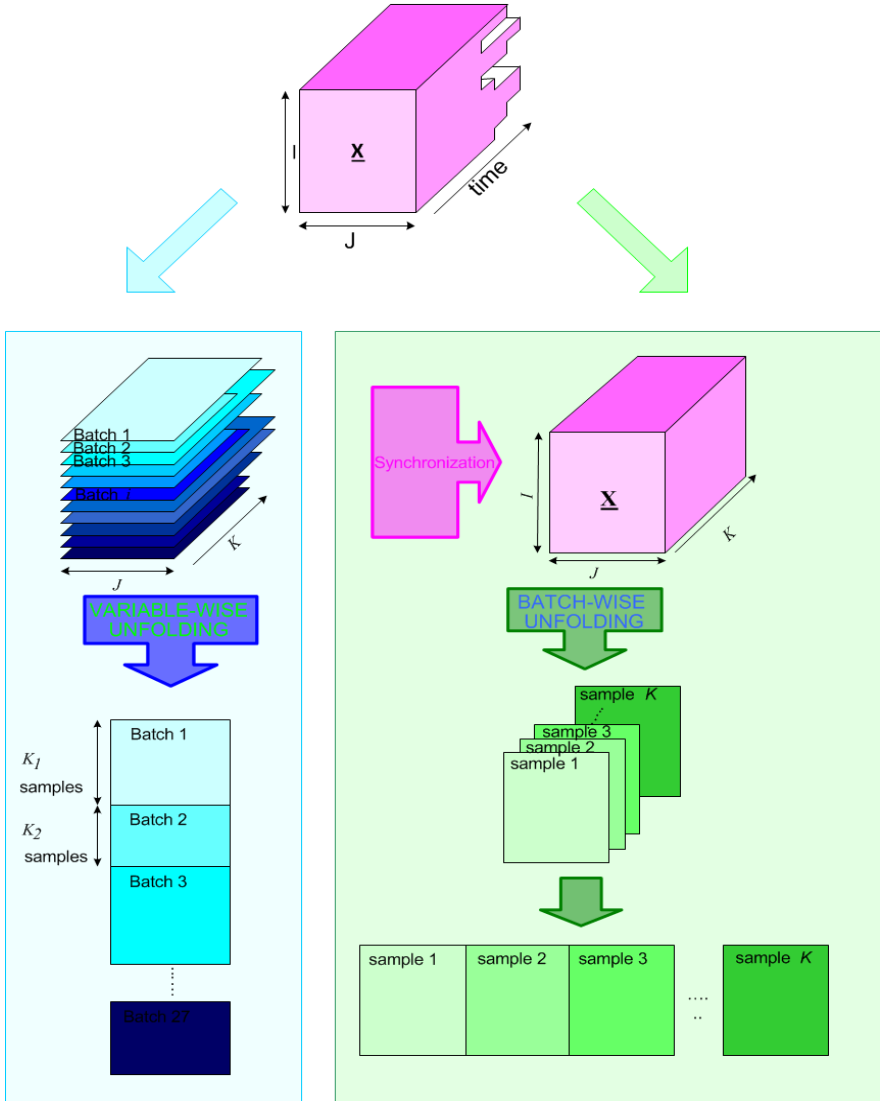


Discussion

- What do you think the main differences are between the two unfolding methodologies?
  - start thinking to the differences in data pretreatment
  - think to the meaning of scores and loadings

Three-dimensional data unfolding

- Multivariate statistical techniques deal with bi-dimensional data
  - to treat 3D arrays the matrix are reorganized in a bi-dimensional fashion
- The most common three-way array data **unfolding** are:
  - **variable-wise unfolding (VWU)**
    - refers to the global mean of the variables
    - **does not consider batch dynamics**
    - accounts for **overall variable correlation**
  - **batch-wise unfolding (BWU)**
    - refers to mean time trajectories of the variables
    - considers **batch dynamics**
    - requires batch **synchronization**
    - accounts for:
      - cross-correlation
      - auto-correlation
      - correlation between variables in time



Example of baker’s yeast

Backer yeast fermentation process

- Process: **yeast fermentation** in Jastbolagest (Sweden)

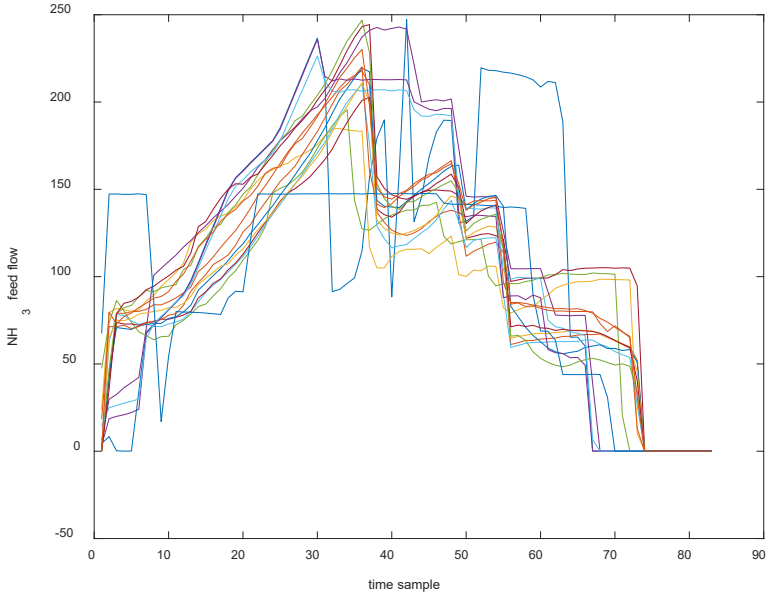
- Available data:
  - 33 batches:
    - 16 reference calibration batches
    - 17 new batches
  - batch duration: 14 hours
  - 83 time samples per batch
  - 7 variables

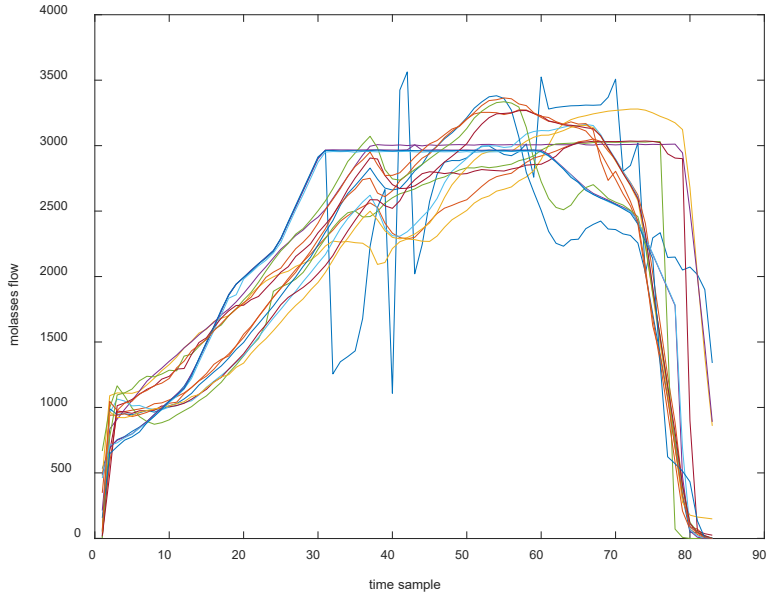
- AIM: process **monitoring**

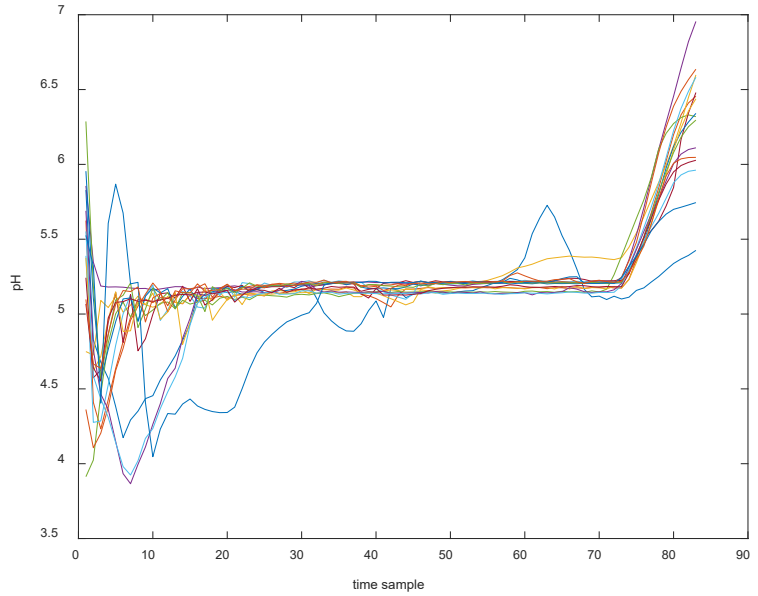
	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

Inspection of the variables time profiles

- A lot of information is stored into the data time profiles:
  - start with **data visualization**, as usual
- Time profiles of the variables show very large variability:
  - scarce automation
  - manual intervention of the operators
  - large variability due to bio-diversity
  - different initial state of the process equipment

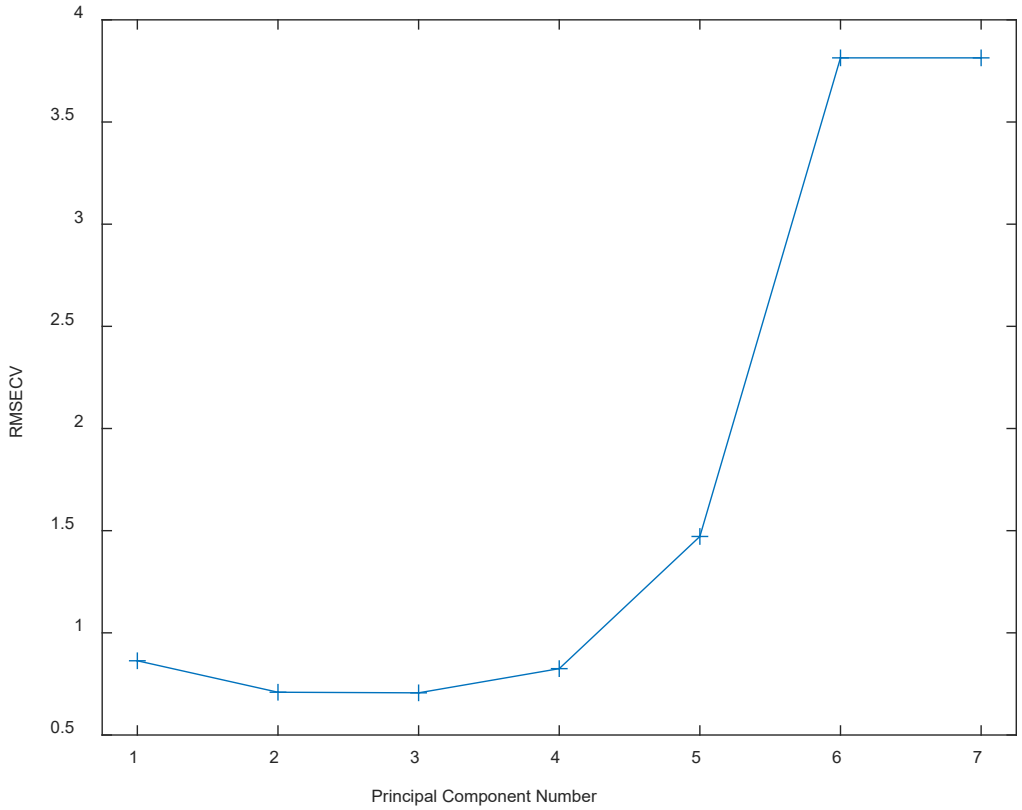






Variable-wise unfolding

Variable-wise unfolding PCA monitoring model



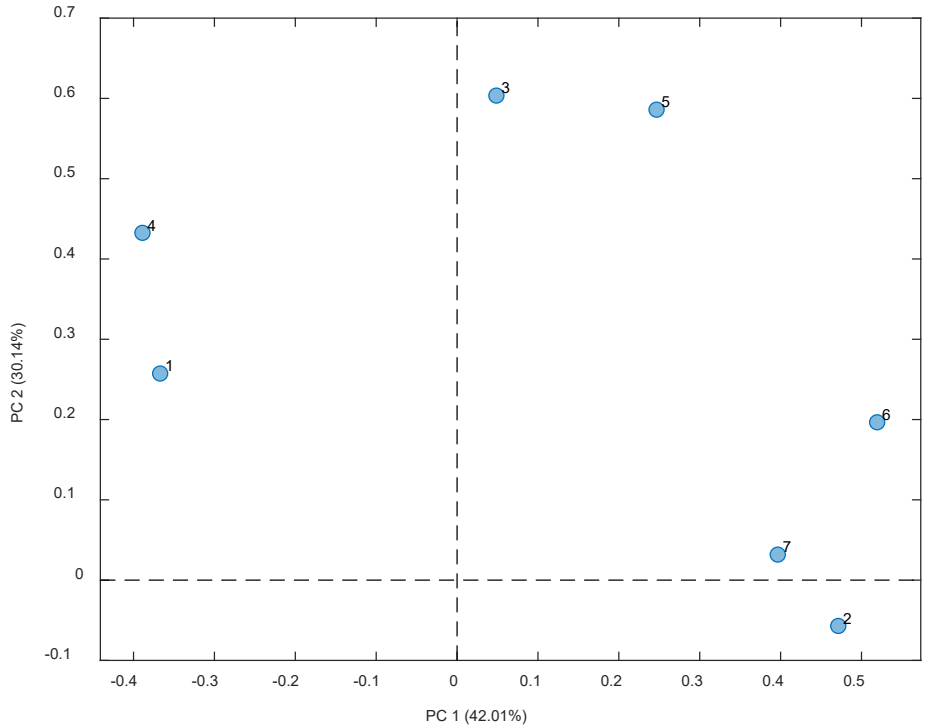
- 16 batches in which 83 time samples of 7 variables are organized in a matrix:

$$\mathbf{X} = [NK \times V] = [16 \cdot 83 \times 7] = [1328 \times 7]$$

tall array!

- 3 PCs are chosen to explain 84.5% of the total variability

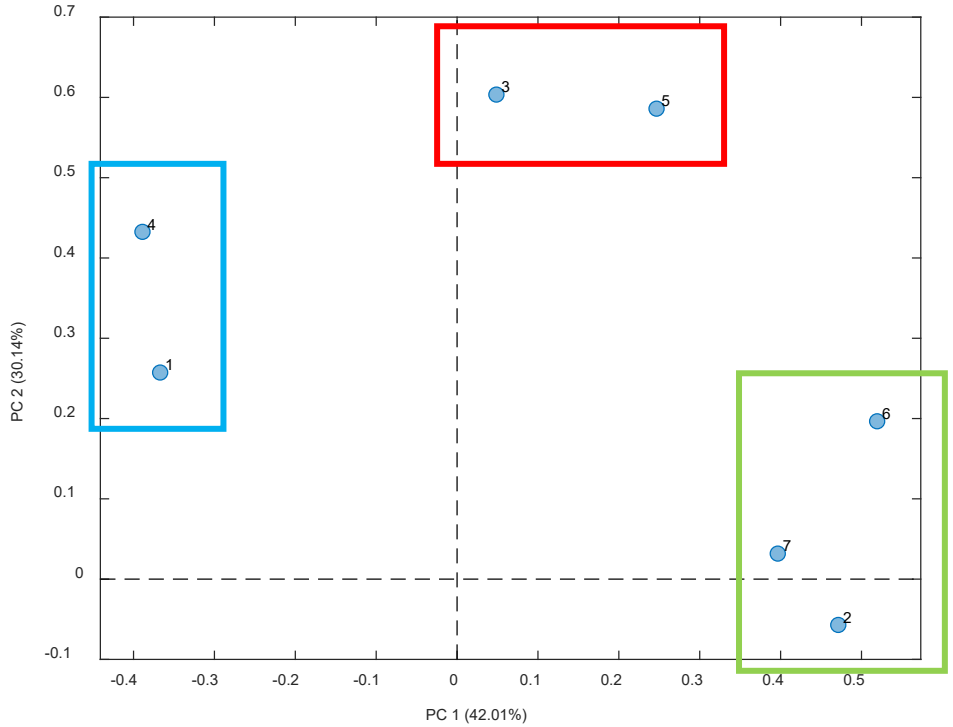
Variable correlation assessed through loadings



	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

Variable correlation

- The **(time-averaged) variable contribution** to PCs are:
  - PC1 (42%) is mainly related to:
    - the positive correlation among temperature, level and pH…
    - NH_3 flow and ethanol are positively correlated
    - temperature, level and pH are negatively correlated to NH_3 flow and ethanol
  - PC2 (30%) mainly related to:
    - positive correlation among molasses flow and air flow, partially correlated to NH_3

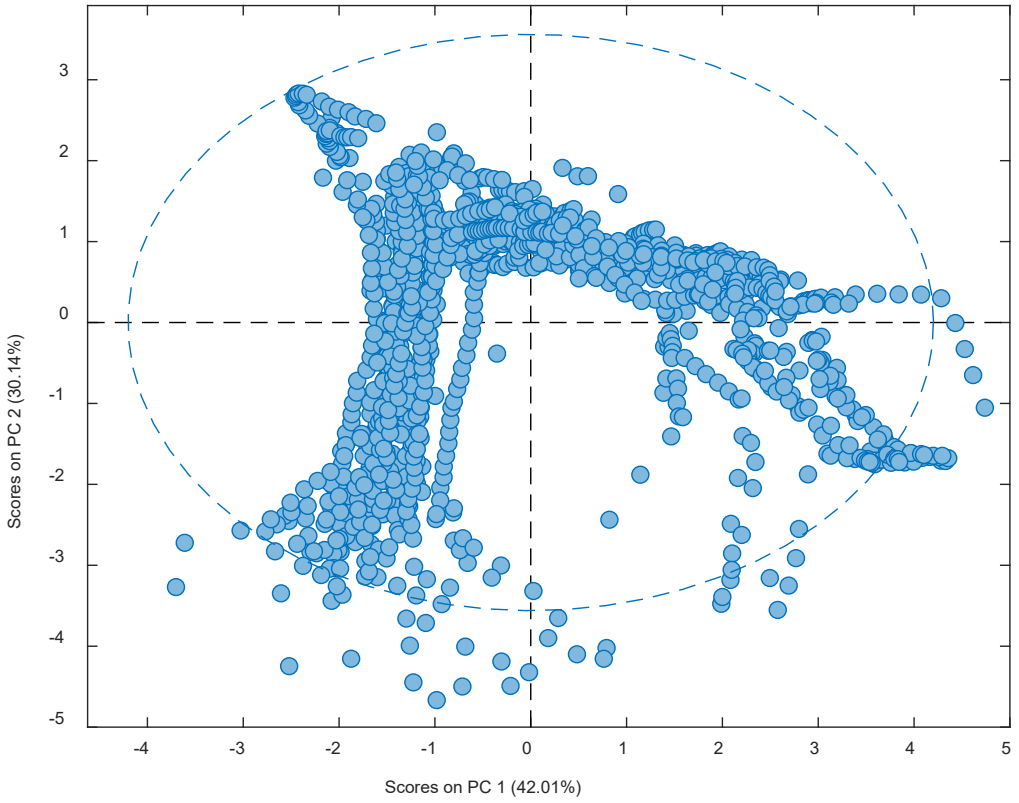


	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

Score plot (1/2)

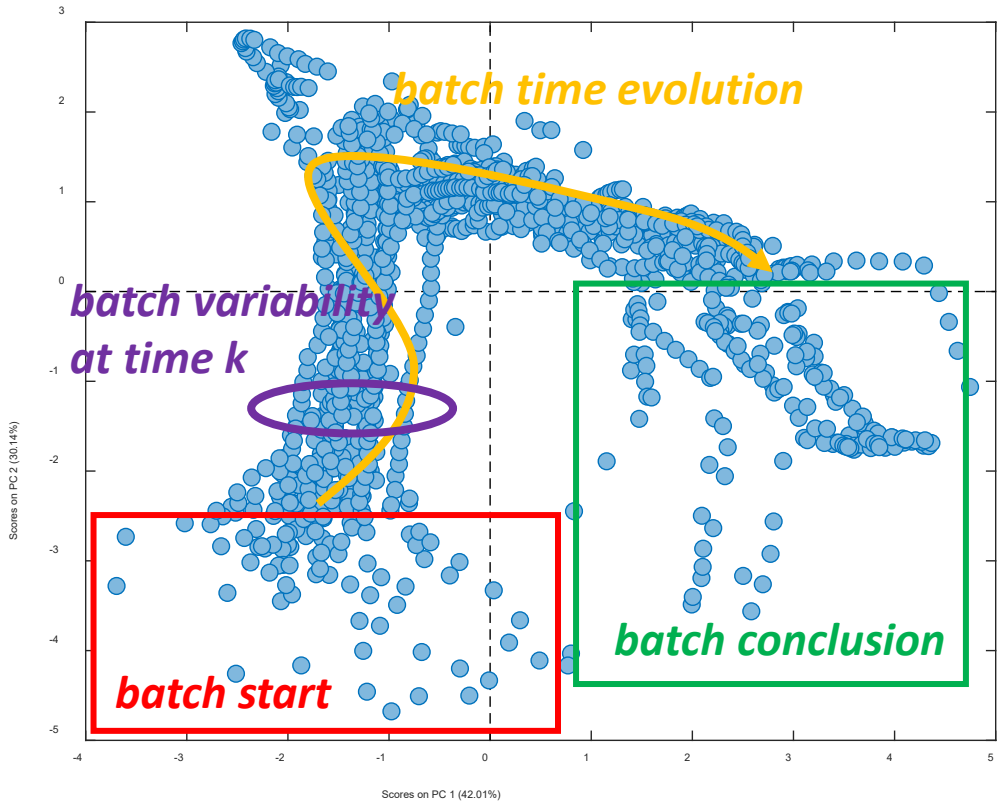
- In the score plot it is possible to observe a peculiar pattern:
  - is the elliptical confidence limit appropriate?

NO! The points are not normally distributed in the score space

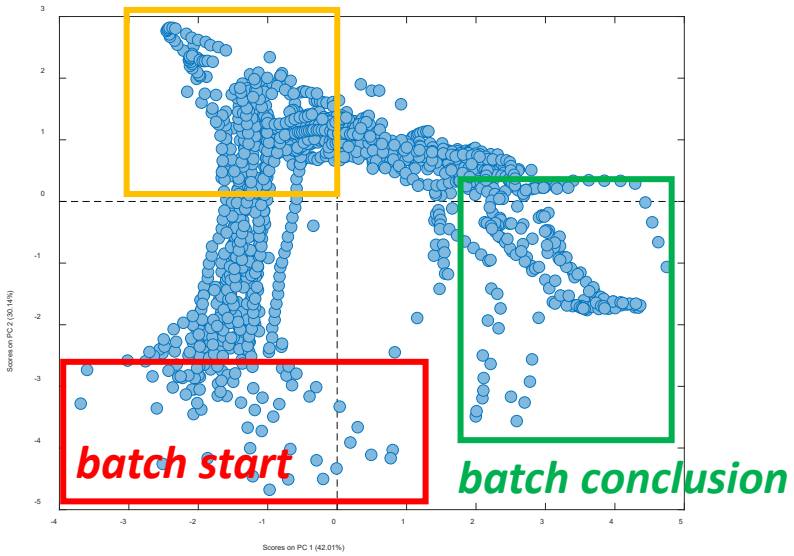


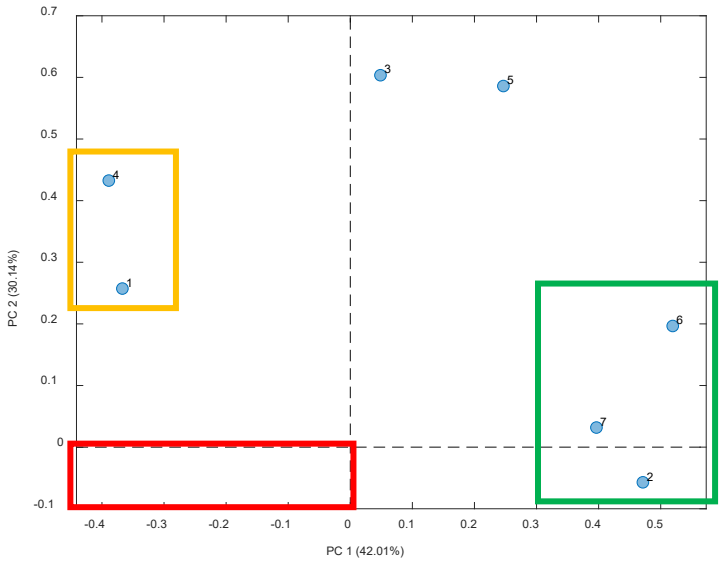
Score plot (2/2)

- In the score plot it is possible to observe a peculiar pattern:
  - a map of the **batch time profile** is available
    - **each point represents all the $V = 7$ variables in a single time point k with $k = 1, \dots, 83$**
  - a batch time trajectory is identifiable
  - zones of batch start and batch completion can be easily identified
  - the operating stages of the batch can be seen



Bi-plot





	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

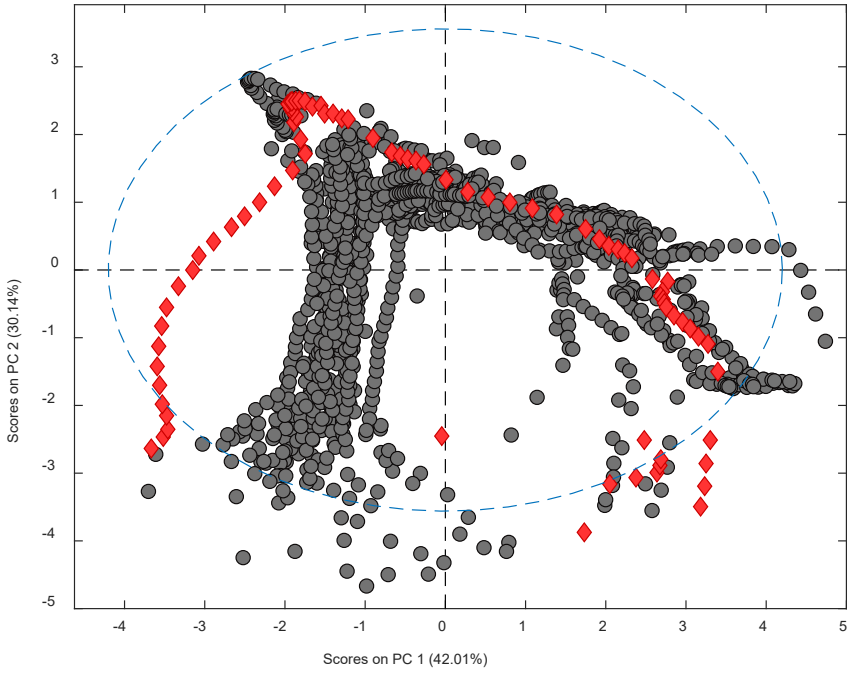
- At the beginning of the batch (red area in both the score and the loading plot):
  - low molasses and air flows
  - low temperature, level and pH

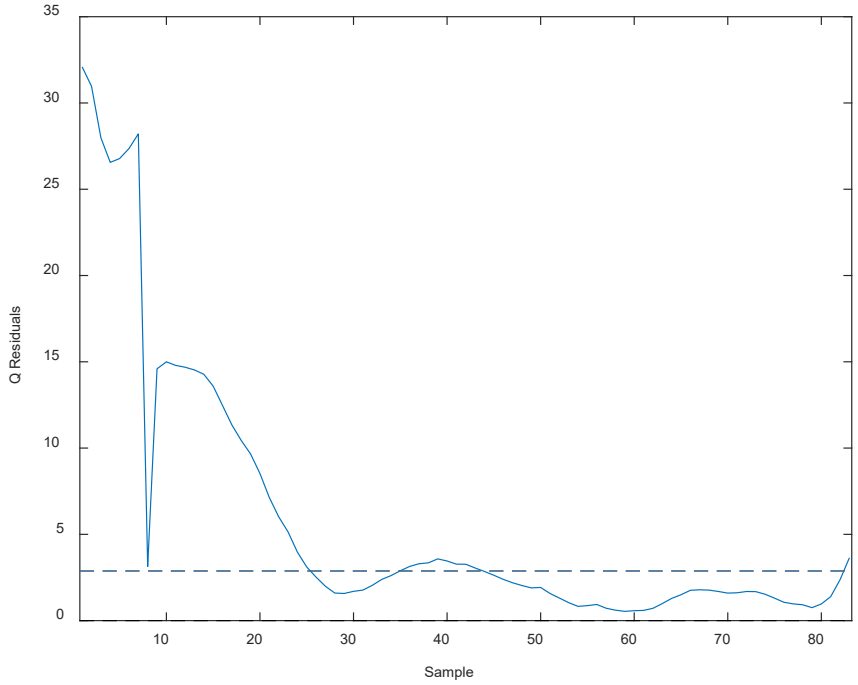
- In the middle of the batch (orange area):
  - high values of molasses flow and air flow
  - high flow of NH_3 and high production of ethanol
  - low values of temperature, level and pH

- The batch conclusion (green area) is characterized by:
  - high values of temperature, level and pH
  - low values of NH_3 flow and ethanol production

Projection of a new regular batch

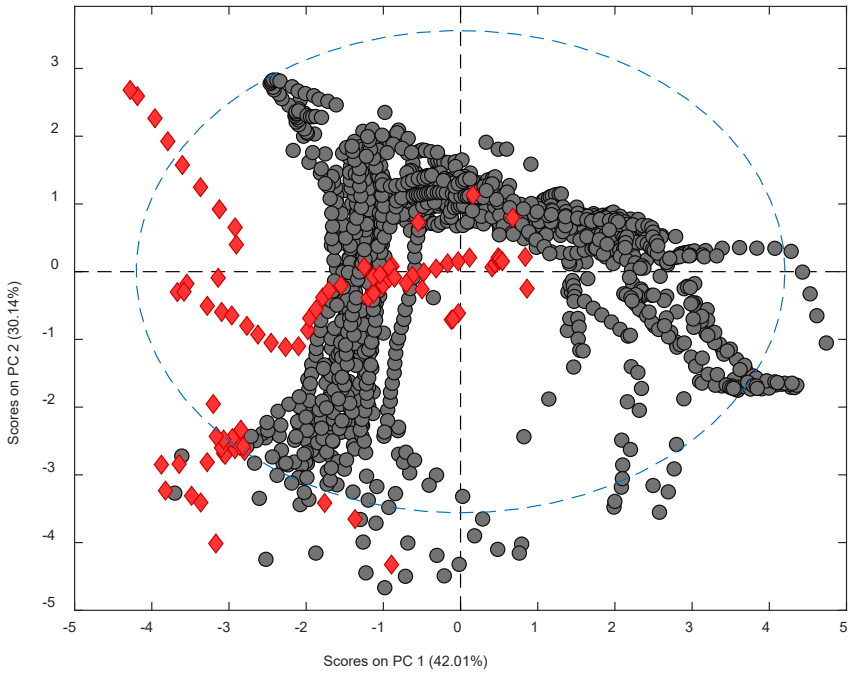
- A new and regular batch is projected onto the PCA model
  - some new features seems to be present in the initial part of the batch
  - a unique PCA model is forced to represent all the operating phases of a batch
    - the variables time profiles for the entire batch are compared to their overall average
      - is it fair for a batch process?
    - the representation of batch start and batch end are not represented in an optimal manner

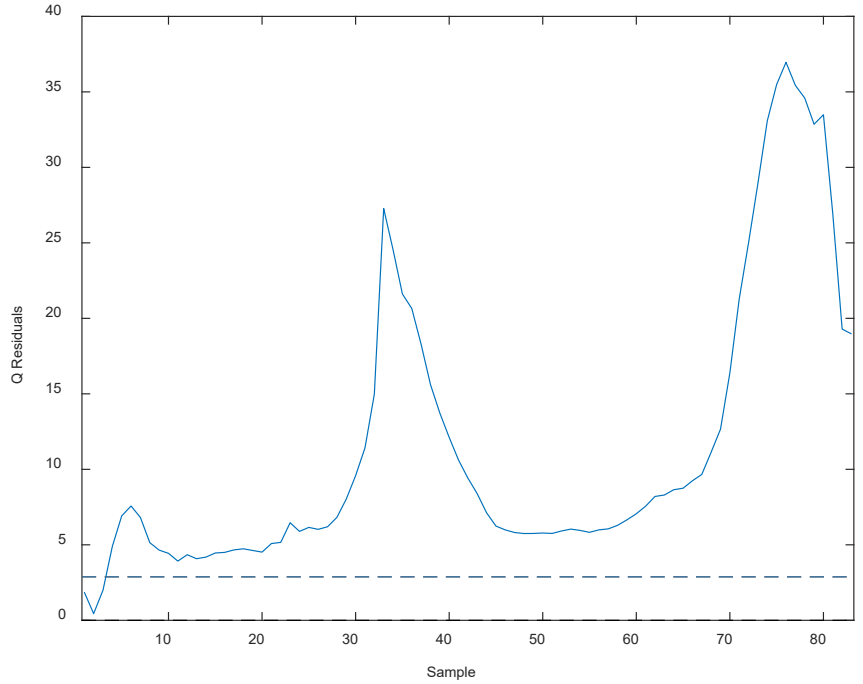




Projection of a new anomalous batch

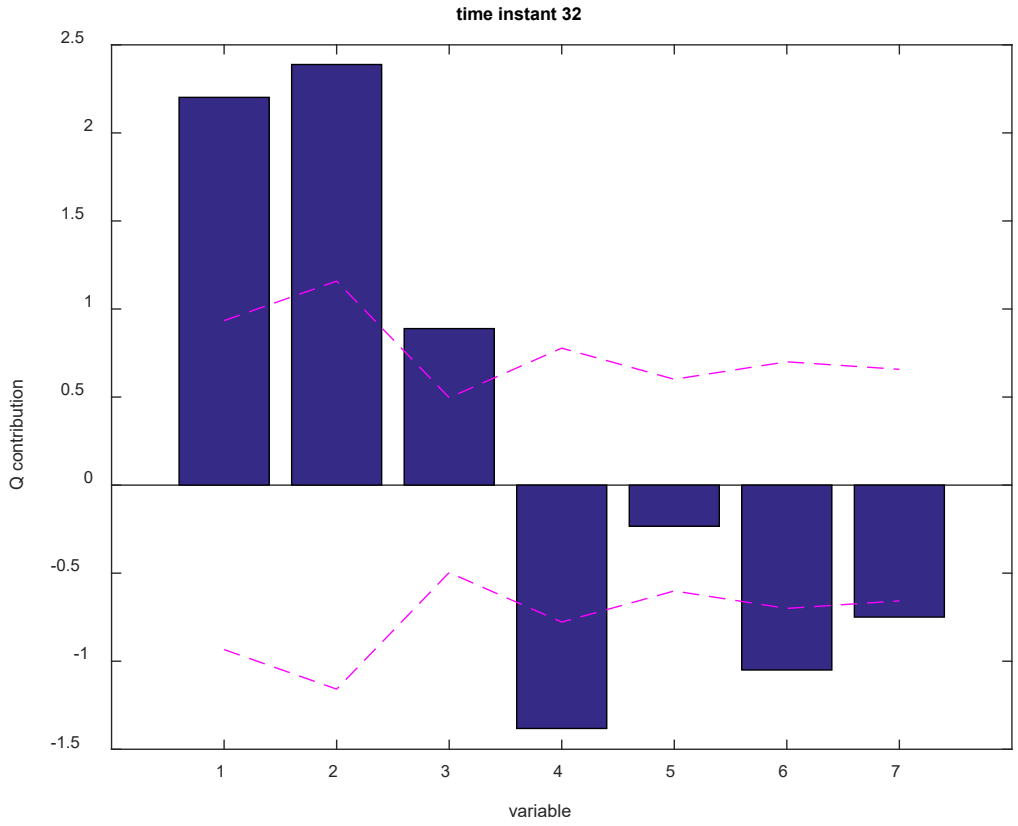
- The projected batch does not follow the typical time trajectory within the score space
- A completely new variable correlation pattern is present in this batch





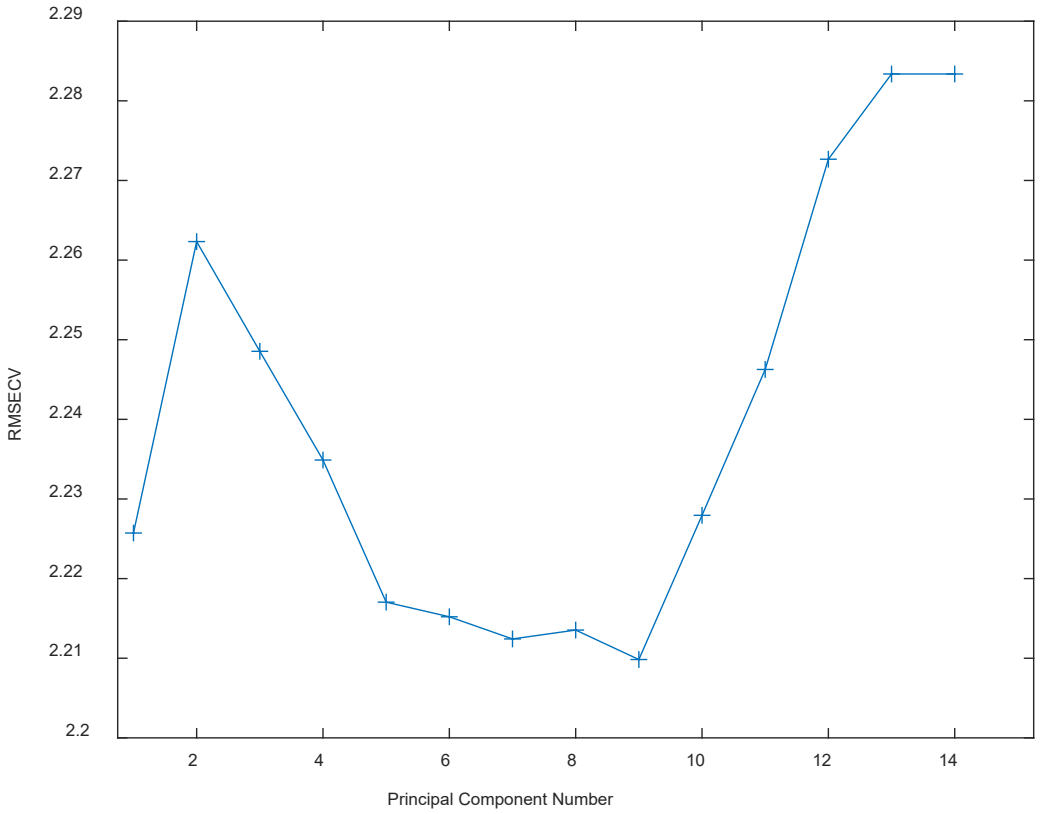
Diagnosis of the fault

- All the variables show a new correlation structure with other ones in the contributions to the residuals
  - here it is shown the contribution plot **at time instant 32**



Batch-wise unfolding

Batch-Wise Unfolding PCA monitoring model



- 16 batches in which 83 time samples of 7 variables are organized in a matrix:

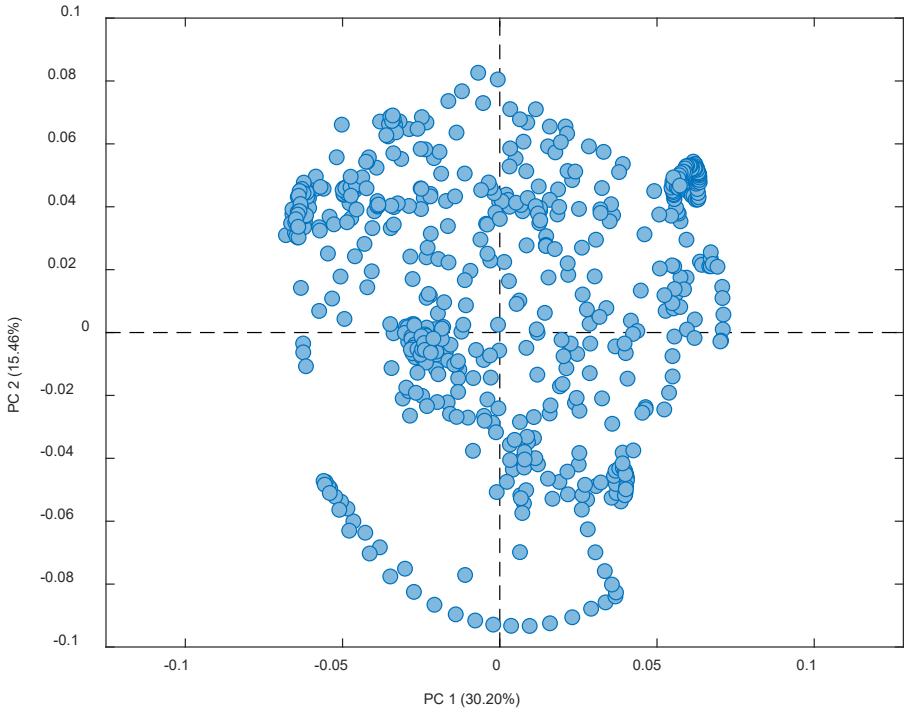
$$\mathbf{X} = [N \times VK] = [16 \times 7 \cdot 83] = [16 \times 581]$$

fat array!

- 9 PCs are chosen explaining 94% of the total variability in the 581 columns (i.e., variables multiplied by time instants)

Variable correlation in the loadings (1/3)

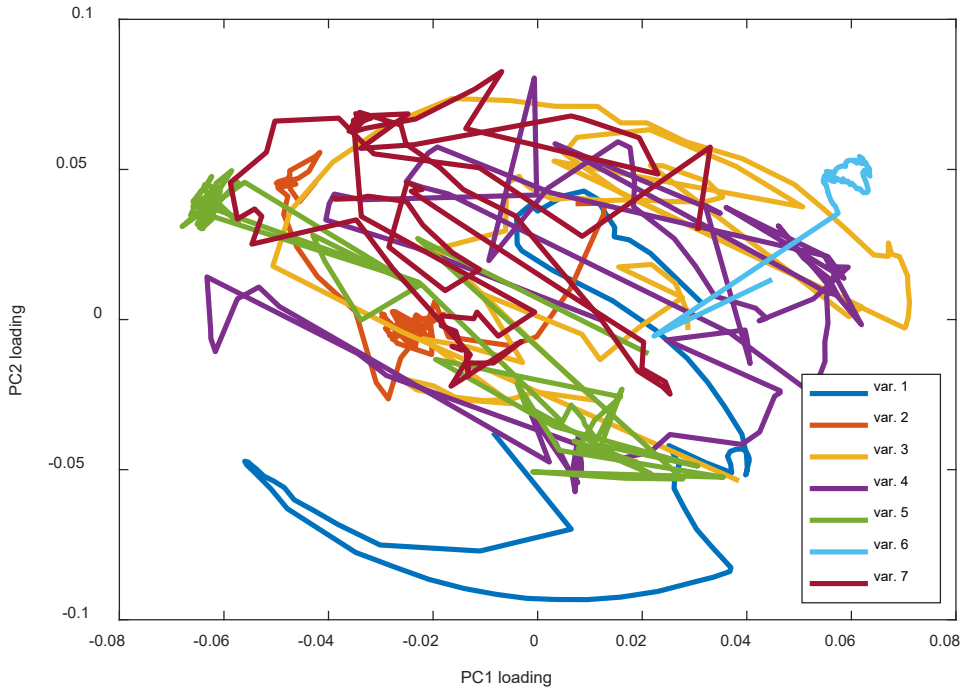
- The loading plot is very complicated to be interpreted



	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

Variable correlation in the loadings (2/3)

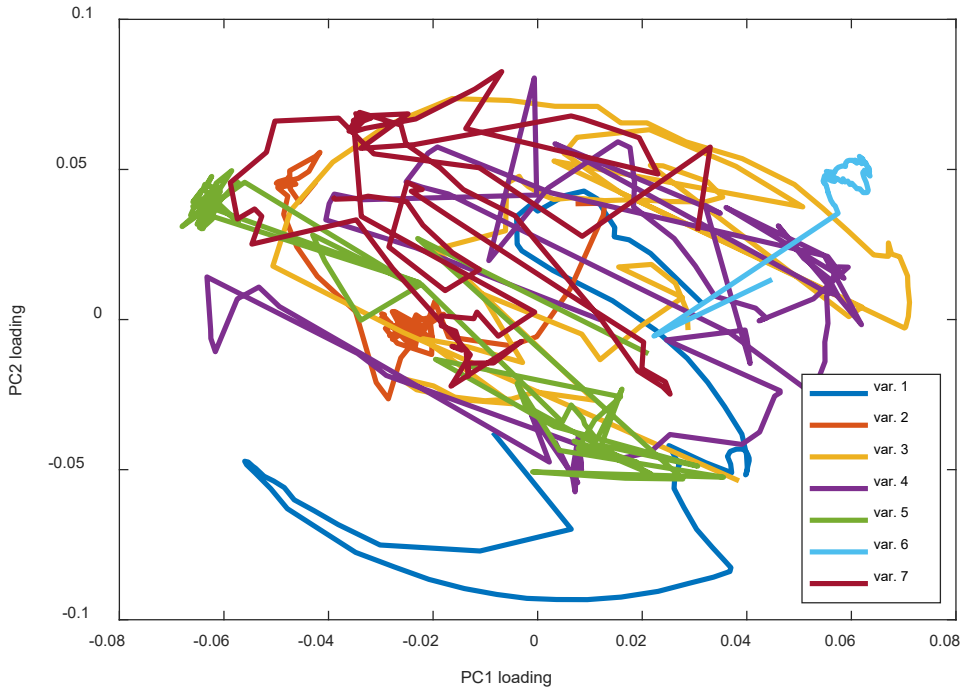
- Consider that in this case the loadings retain information on time:
  - loadings time trajectories must be considered
  - however, the interpretation is often not straightforward…



	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

Variable correlation in the loadings (3/3)

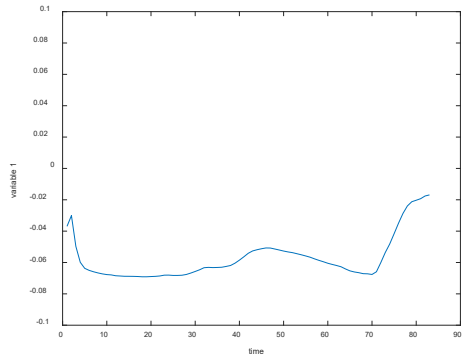
- The loading plot is very complicated to be interpreted:
  - the **variables time trajectories are present in the loading plot**
    - the importance of the variables at each time instant can be observed
  - PC1 is mainly explained by the NH_3 flow and the molasses flow, but anti-correlated to the air flow in the central part of the batch
  - PC2 is mainly related to pH and molasses flow, which is anti-correlated to ethanol

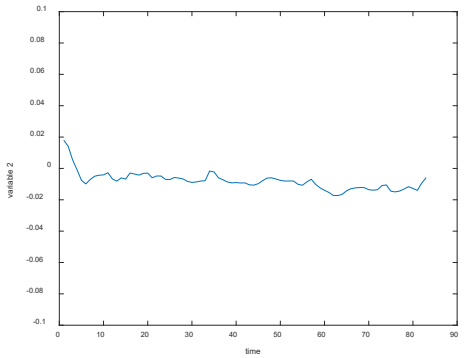


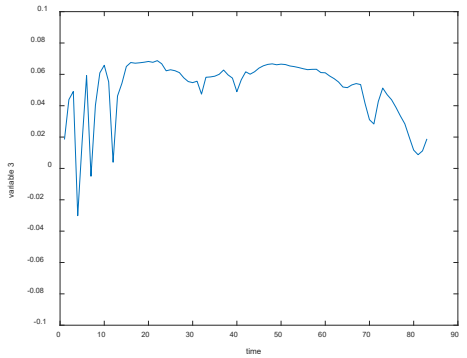
	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

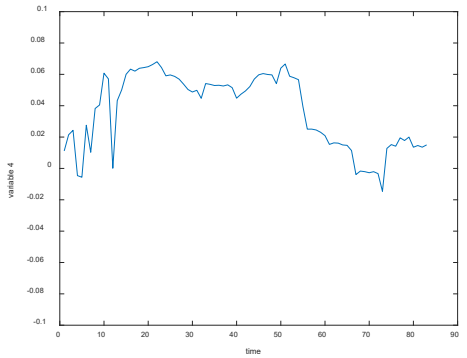
Variable correlation in loading time trajectories

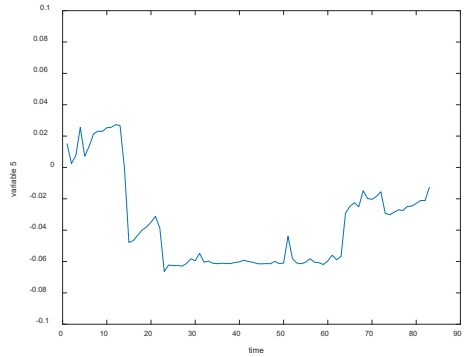
- The **time trajectories of the PC1 loadings** (but also PC2 and PC3 loadings) for each variable are studied and determine the importance of the variables in time
  - molasses flow (variable #3) and NH3 flow (#4) in the middle of the batch, and level (#6) at the end of the batch are positively correlated and their high levels are represented in positive PC1
  - ethanol (#1) and air flow (#5) are positively correlated, but anti-correlated with the previous ones and their high levels are represented in negative PC1
  - temperature (#2) and pH (#7) seems to be less important (i.e., close to zero) for PC1 variability

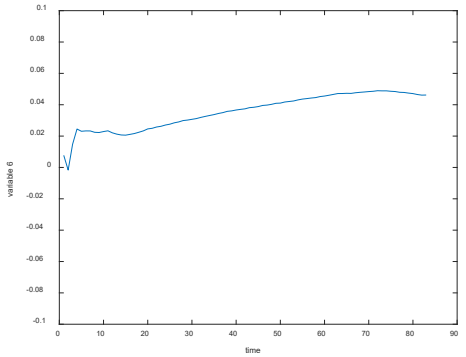


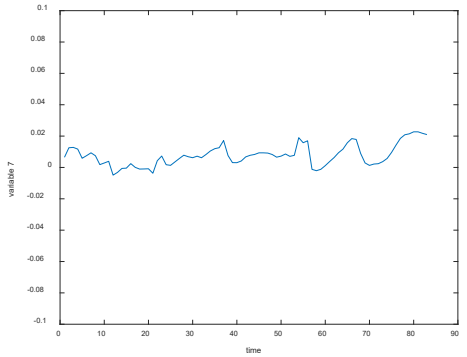








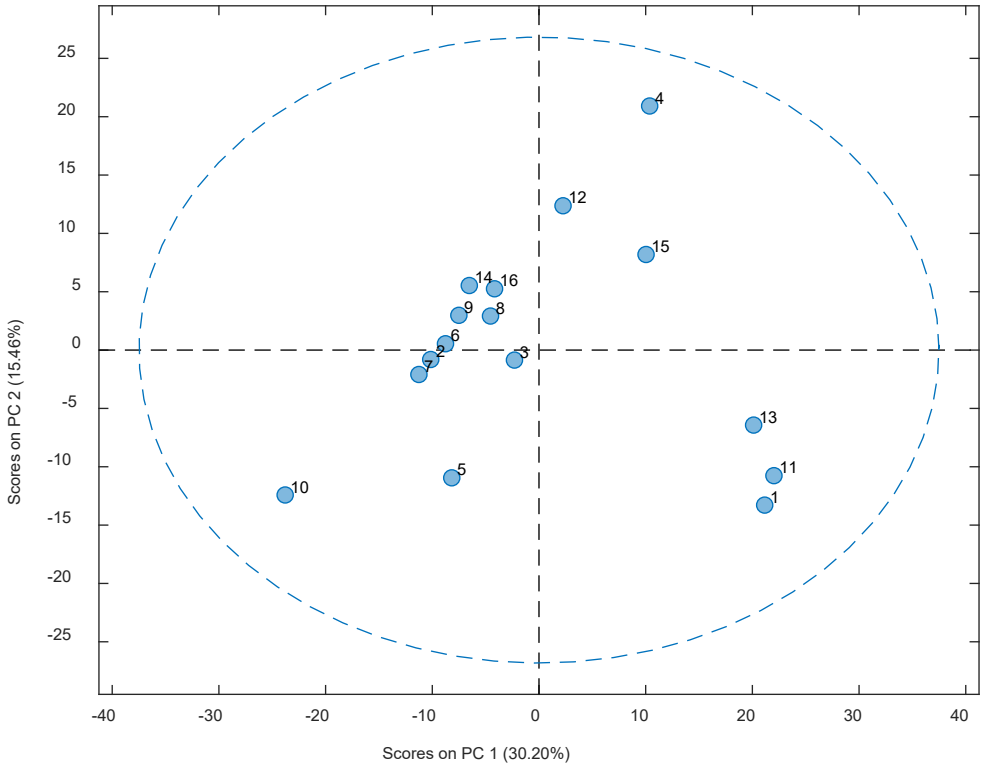




	variables
1	ethanol
2	temperature
3	molasses flow
4	NH_3 flow
5	air flow
6	level
7	pH

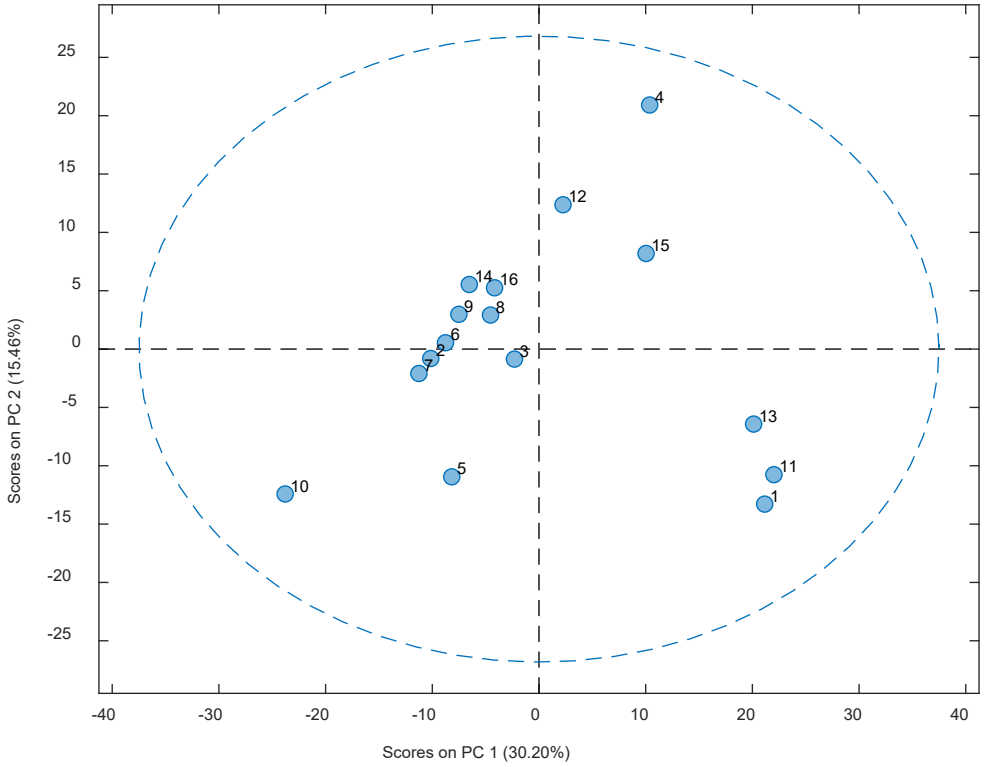
Score plot (1/2)

- What is the information stored in the score plot of a batch-wise unfolded data PCA model?



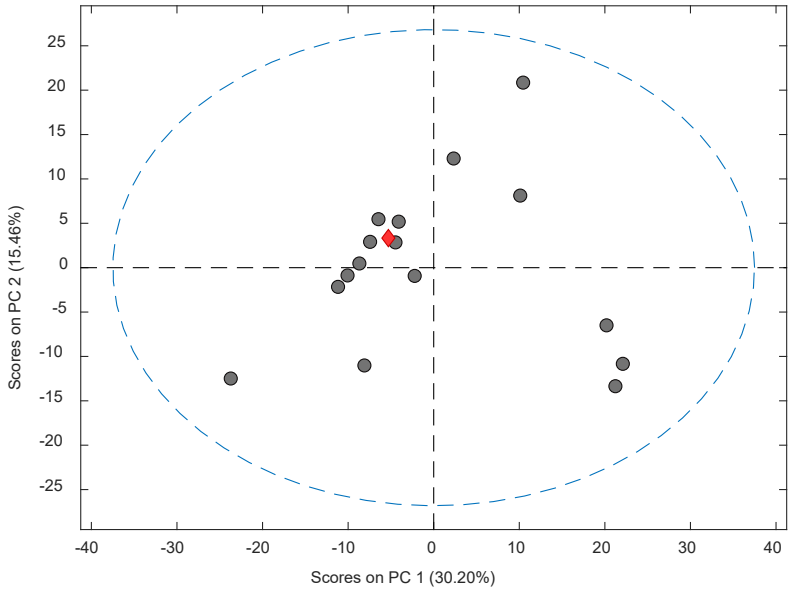
Score plot (2/2)

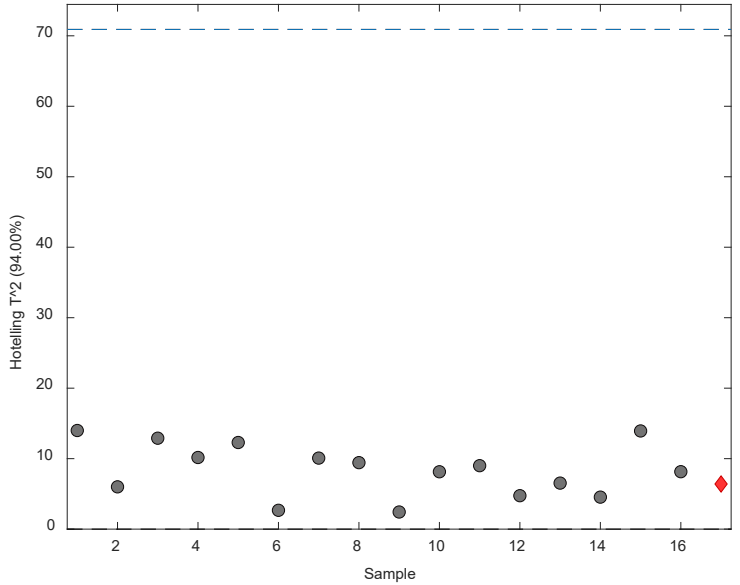
- In the score plot it is possible to observe:
  - how **batches** are related
    - **each point represent one batch** (i.e., all the variables time profiles of one batch for the entire batch duration)
    - the monitoring chart can judge the status of an **entire completed batch from the variables time profiles**

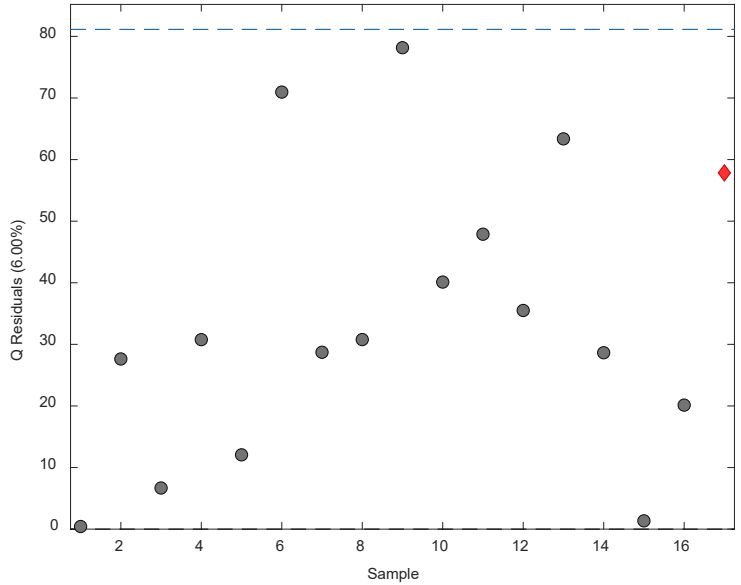


Projection of a new regular batch

- A **regular batch** is projected after batch completion
  - it is detected to be a completely regular batch in:
    - score space
    - T^2 monitoring chart
    - Q monitoring chart

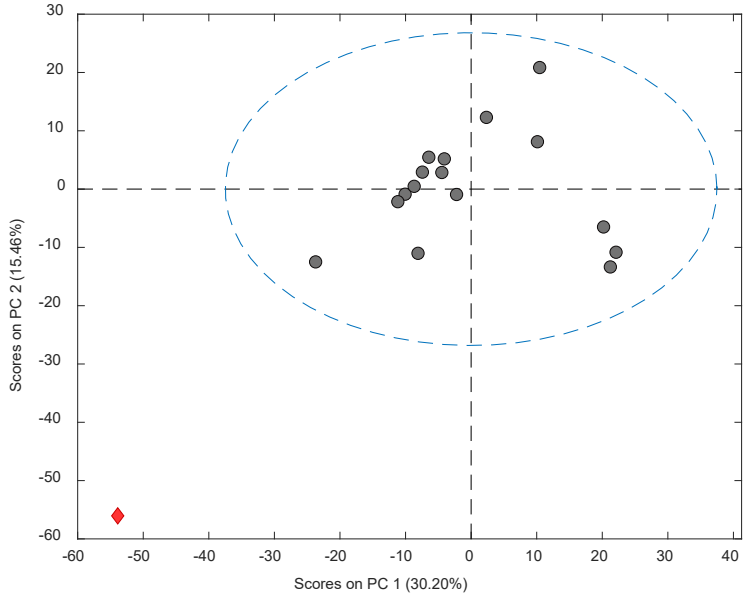


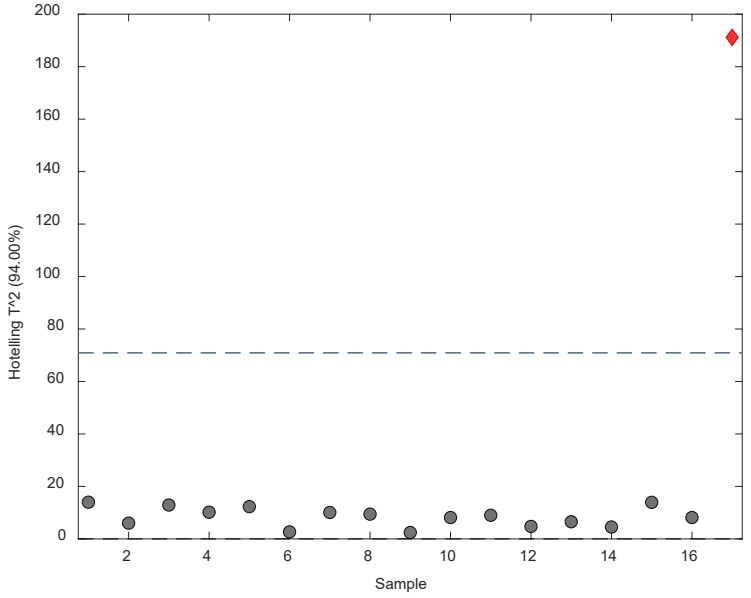


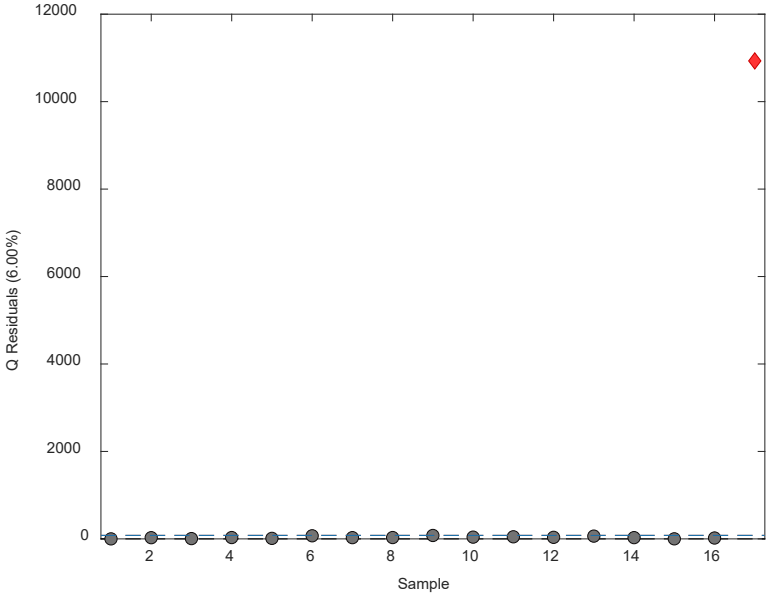


Projection of a new anomalous batch

- The **anomalous batch** is out of the 95% confidence limits in both:
  - Hotelling T^2 chart
  - the Q residual chart

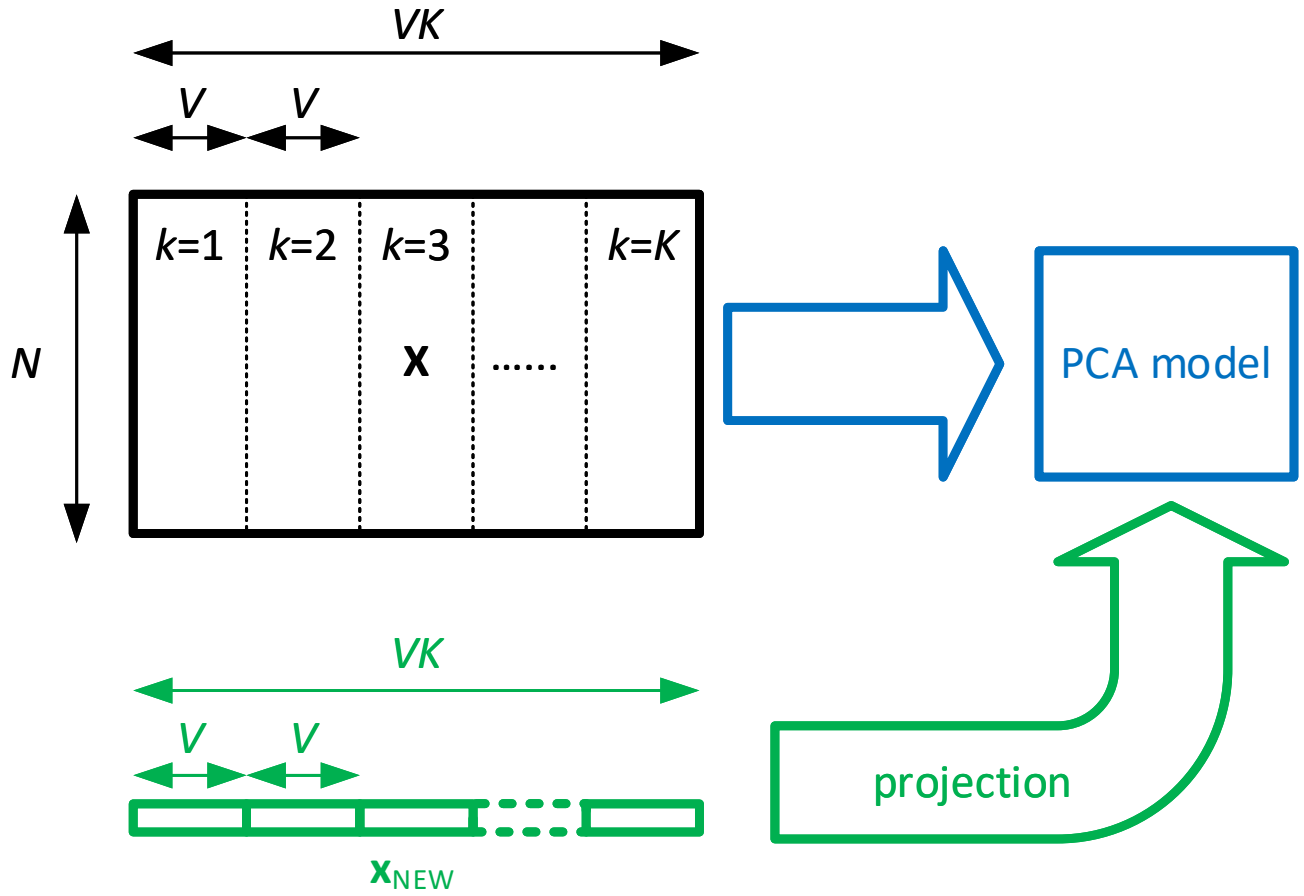






New completed batch projection on PCA

- After batch completion



Post-mortem monitoring

- Batch-wise unfolding monitoring can be carried out only if the measurement for the **entire batch** are available
  - **post-mortem monitoring**
- Do you think there could be more desirable ways to monitor a process?

Post-mortem monitoring vs. online monitoring

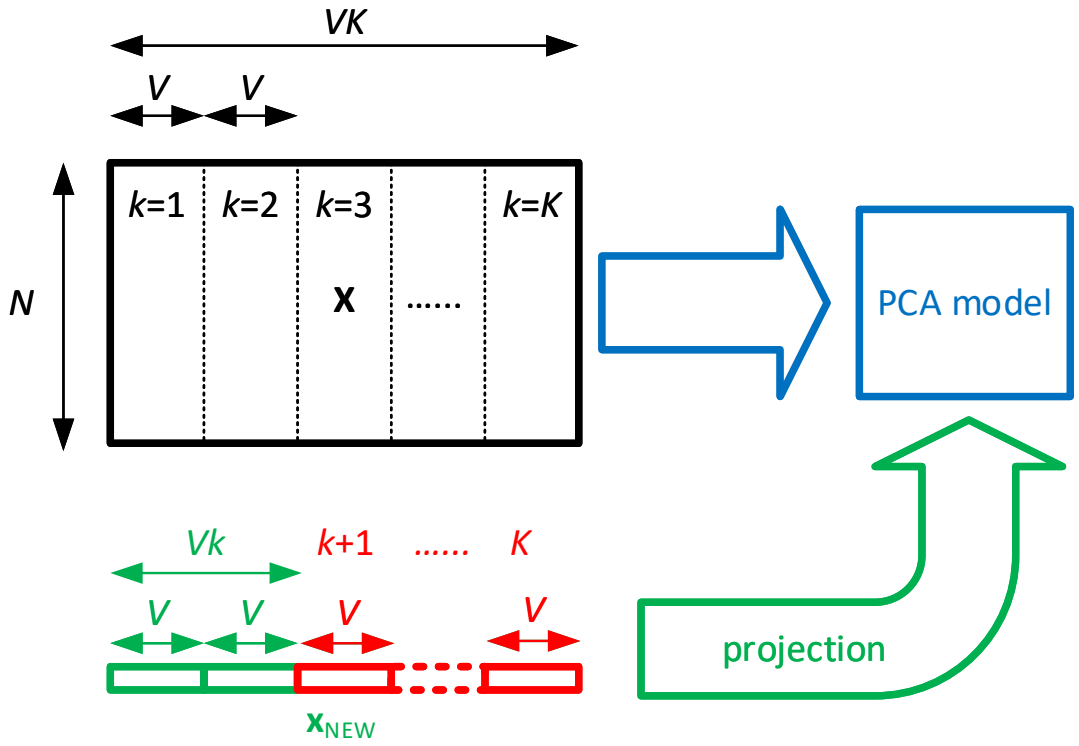
- Batch-wise unfolding monitoring can be carried out only if the measurement for the entire batch are available
  - post-mortem monitoring
- … however, the process monitoring can be carried out also **in real time**:
  - a strategy is needed for batch completion
  - what do you think the best strategy for batch completion could be?

Post-mortem monitoring vs. online monitoring

- Batch-wise unfolding monitoring can be carried out only if the measurement for the entire batch are available
  - post-mortem monitoring
- … however, the process monitoring can be carried out also **in real time**:
  - a strategy is needed for batch completion
  - batch completions is performed at time instant k by considering that:
    - the future missing part of the batch (time instants $k+1$ to K) is filled artificially
    - batch completion at time instant k can be done considering 2 alternative strategies:
      1. the batch at its average conditions for future time instants from $k+1$ to K
      2. the current (i.e., at time k) deviation from the average is maintained constant for the rest of the batch

New batch projection on PCA in real time

- At each time $k = 1, 2, \dots, K$
  - V variables available until time instant k
  - missing future measurements from time $k+1$ to K

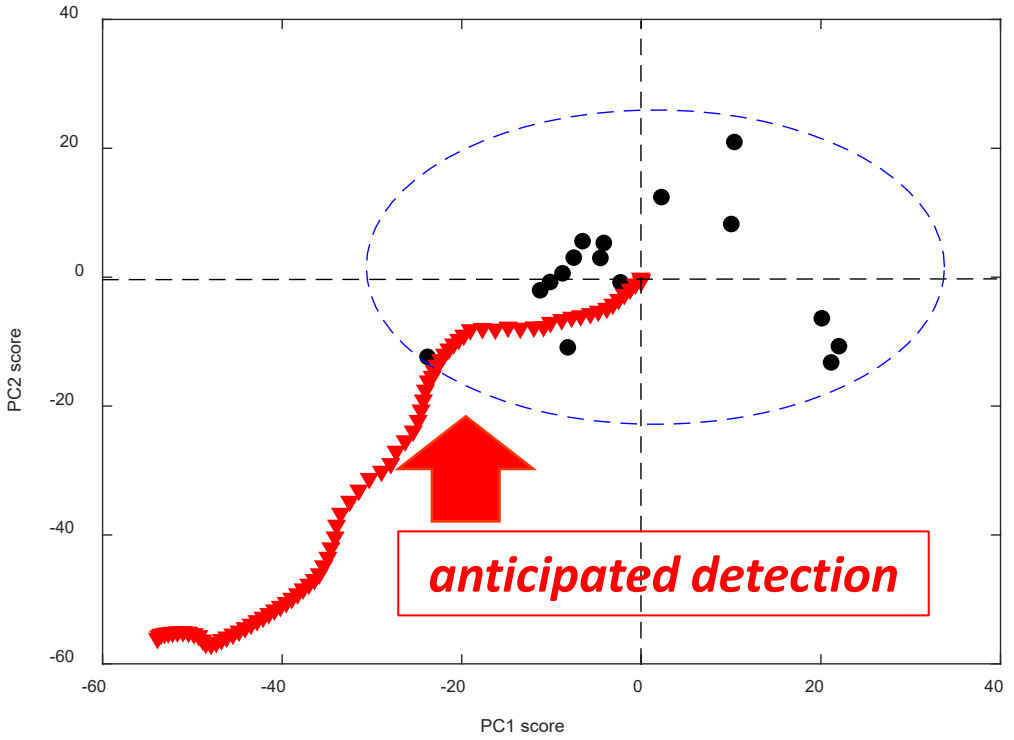


Alternative solutions:

1. fill with zeros the autoscaled version of $\mathbf{x}_{NEW}$ for the remaining part of the batch (time samples $k+1$ to K)
2. repeat the k-th autoscaled vector of V variables for the remaining part of the batch (time samples $k+1$ to K)

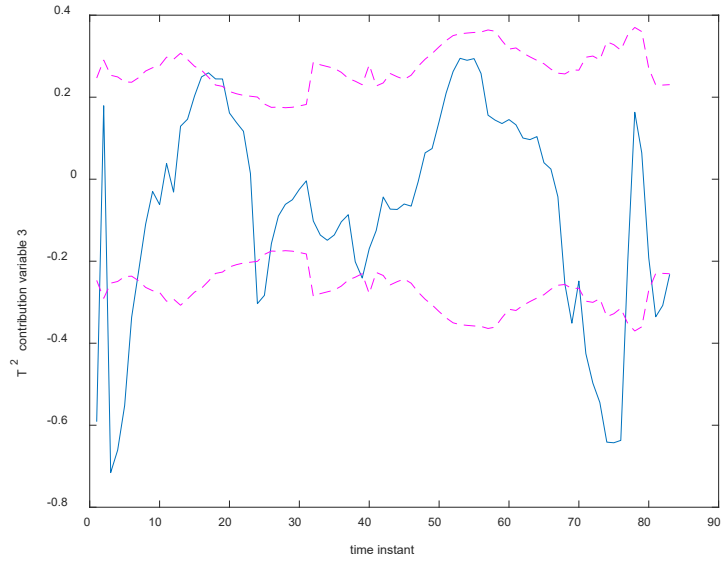
Online process monitoring

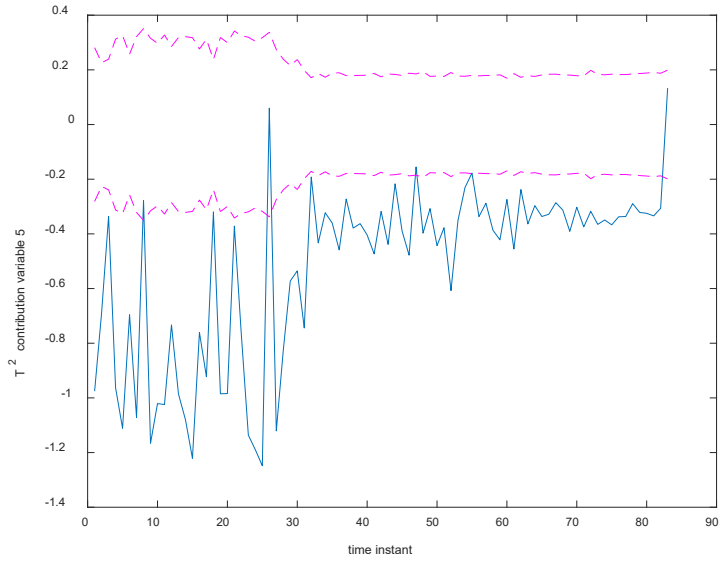
- The monitoring can be performed in real time, with improved detection capability
  - once the measurements are available at time instant k the model is updated
  - the future missing measurements (from time $k+1$ to time K) can be inputted:
    - maintaining the batch at the average conditions
    - maintaining unchanged the deviation from the average conditions

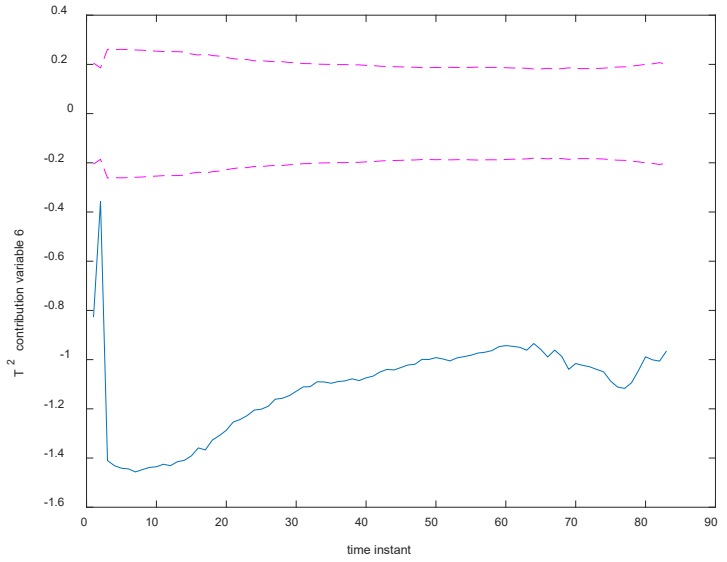


Diagnosis of the fault

- The diagnosis can be done for all the variables time profiles through **time profiles of the contributions' plots** built for all the variables time profiles:
  - molasses flow is at regular values for most of the batch duration
    - low values are experienced at the very beginning and at around time instant 75
  - air flow and level are too low for the entire batch duration







Take-home message

- PCA is commonly utilized also to **monitor batch processes**
  - it treats **multi-way (i.e., multidimensional) datasets**
  - multi-dimensional datasets should be **unfolded**
  - the time dimension is a key do deal with batch data
  - monitoring charts must consider time
  - monitoring can be performed:
    - for the **end-point**, after batch completion
    - **in real time** (which requires batch synchronization)

Today's homework

- Practice with Matlab® and PLS_Toolbox®
  - datasets are available in Moodle for both the examples of:
    - Yeast fermentation

- See the video on the computational **Laboratory #3** in Moodle
  - you will learn how to build monitoring models on **batch processes**
  - you will practice with a new dataset in the case of:
    - monitoring of a batch rubber production

… per sempre a fianco a me!