NTNU

Norwegian University of
Science and Technology

Time dependent data and real-time prediction

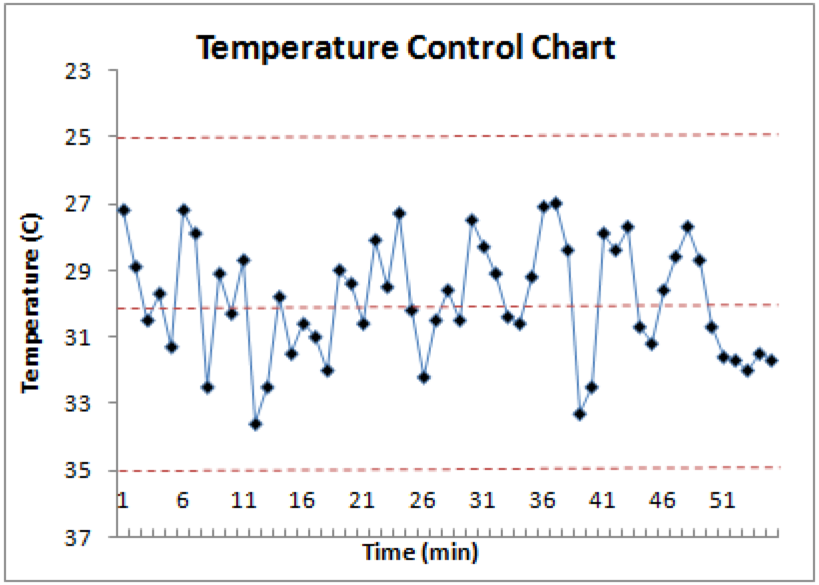
Frank Westad
Department of Engineering Cybernetics
NTNU, Trondheim, Norway
frank.westad@ntnu.no

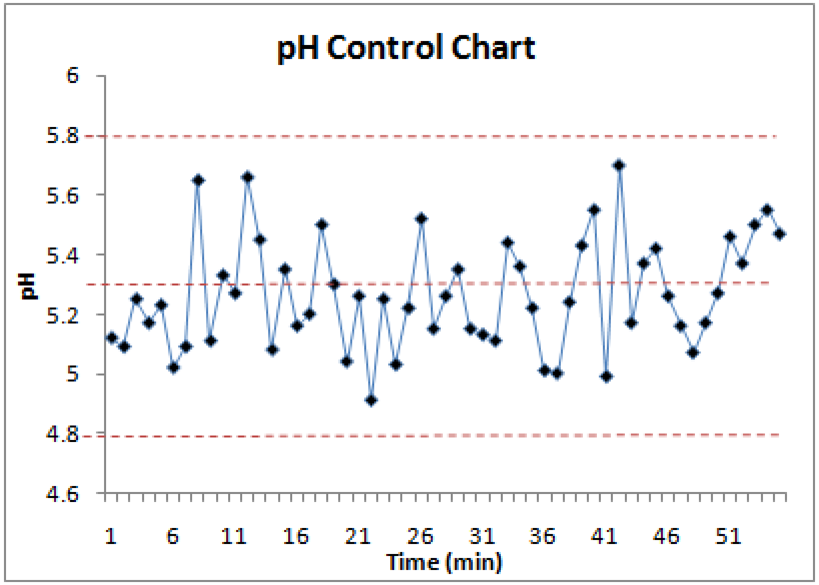
Contents

- Introduction
- Short reminder: Outliers in PCA and multivariate regression
- Projection and prediction in real-time
- Modelling of time-dependent processes
- Examples

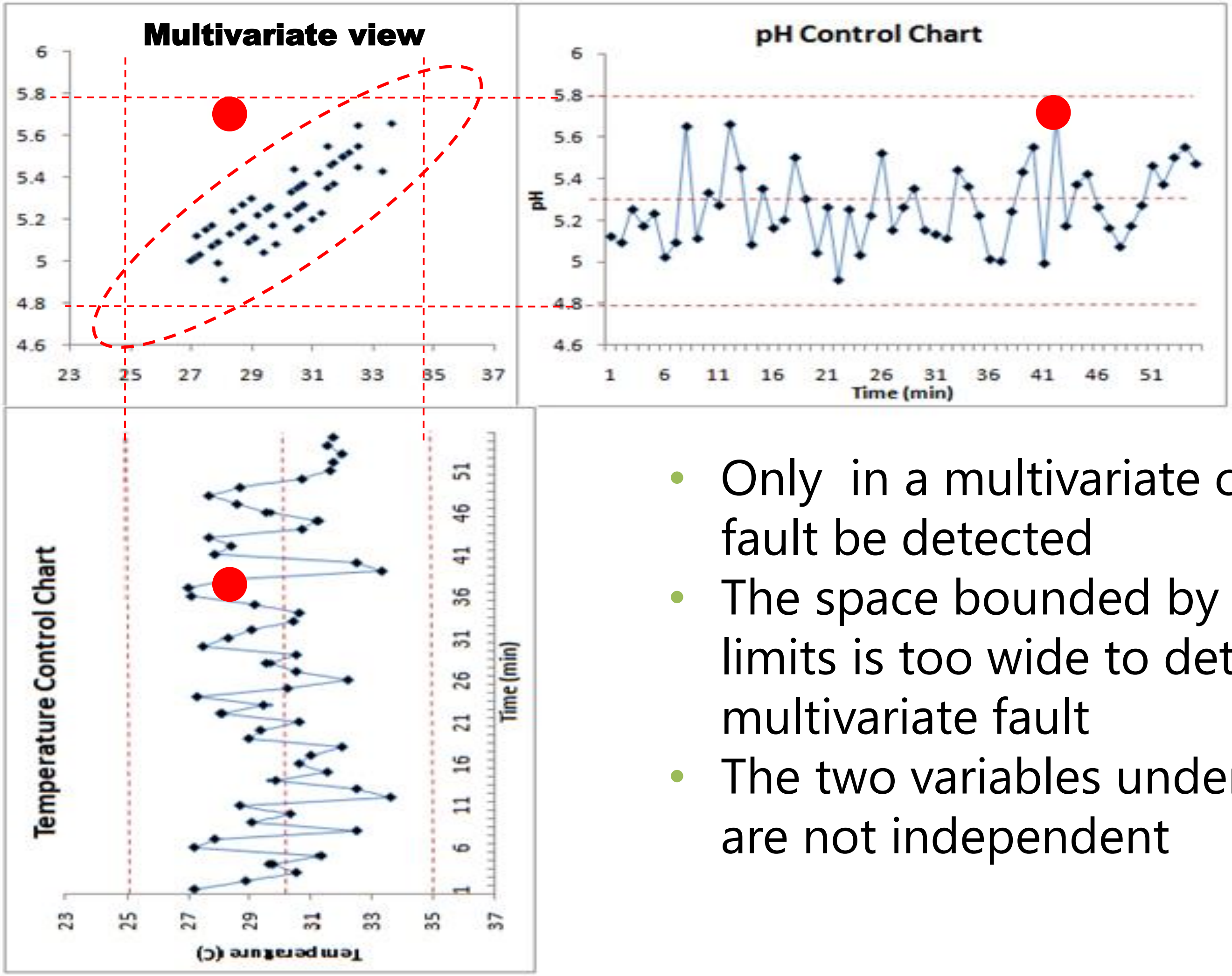
The univariate world

Consider the two plots below, there seems to be no “out of control” situation in the process.



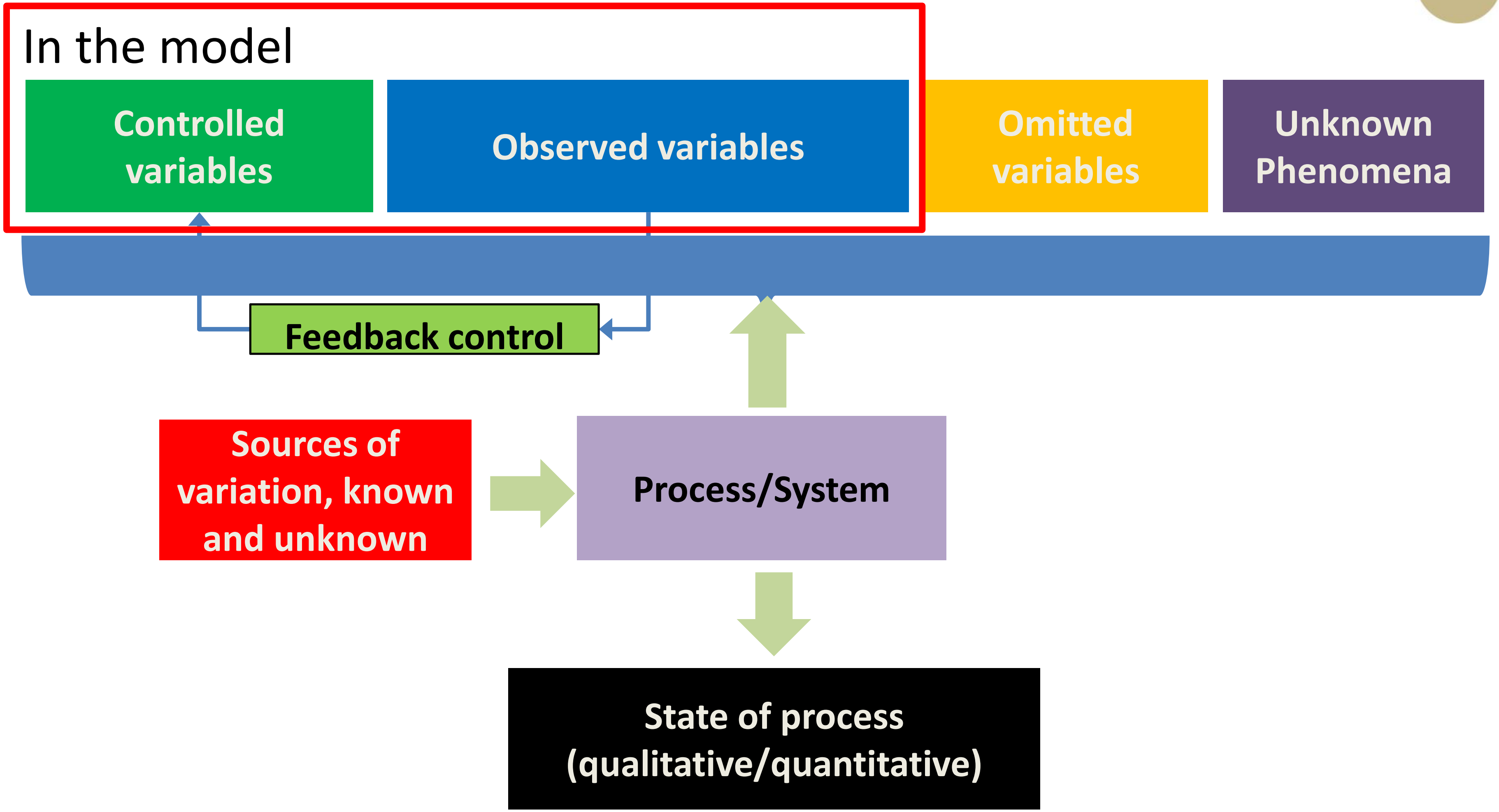


Why we need the multivariate world



- Only in a multivariate context can the fault be detected
- The space bounded by the univariate limits is too wide to detect a multivariate fault
- The two variables under consideration are not independent

Which variables are present in a process/system?



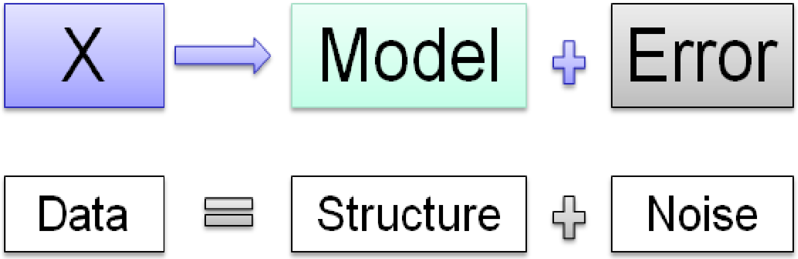
Process modelling

- Types of processes
  - Continuous
  - Batch
  - Semi-batch/semi-continuous
- Some aspects to consider:
  - Is the process stationary?
  - Is there a control loop in operation?
  - Can I use time* actively to improve my predictions?
  - What is the proper validation scheme?

* Time may have various meanings in this context

Principal Component Analysis (reminder)

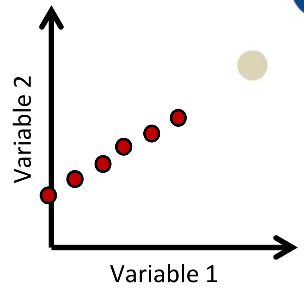
- A method to reduce dimensionality
  - Replace original variables with latent variables
  - Linear combination of the original ones
  - Do not necessarily represent physical factors
  - Useful if variables are correlated
- Information is extracted and noise removed
- Many application areas
  - Exploratory data analysis
  - Classification and identification
  - Variable reduction
  - **Process monitoring**
  - Visual analysis of variance



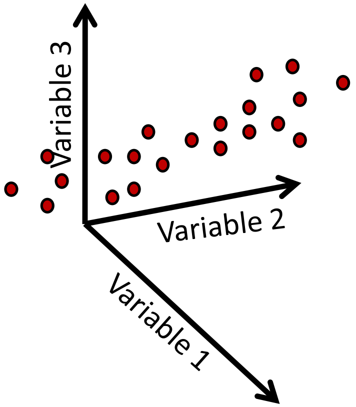
X → Model + Error

Data = Structure + Noise

2D



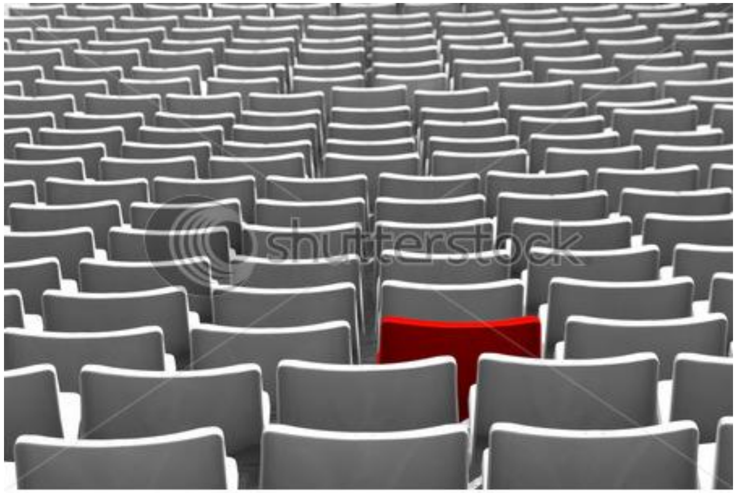
3D



> 3D PCA

What is an outlier? (1/2)

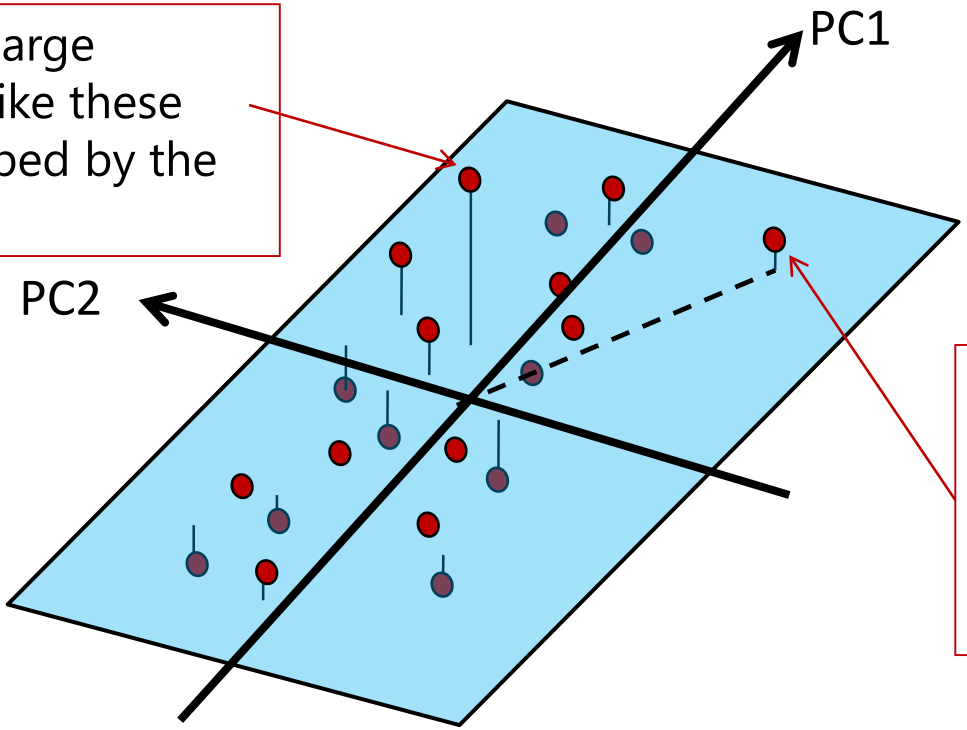
- An outlier is an object which deviates from the other objects in a model and may not belong to the same population as the majority
- Outliers can disturb the model
- The cause of outliers could be one or more of the following:
  - Measurement error
  - Wrong labelling
  - Deviating products / processes
  - Noise
  - Extreme / interesting sample



What is an outlier? (2/2)

This example shows a projection of a set of samples onto PC1 and PC2.

- Residual: Distance to (multidimensional) model plane.
- Leverage: Distance from center along (multidimensional) model plane

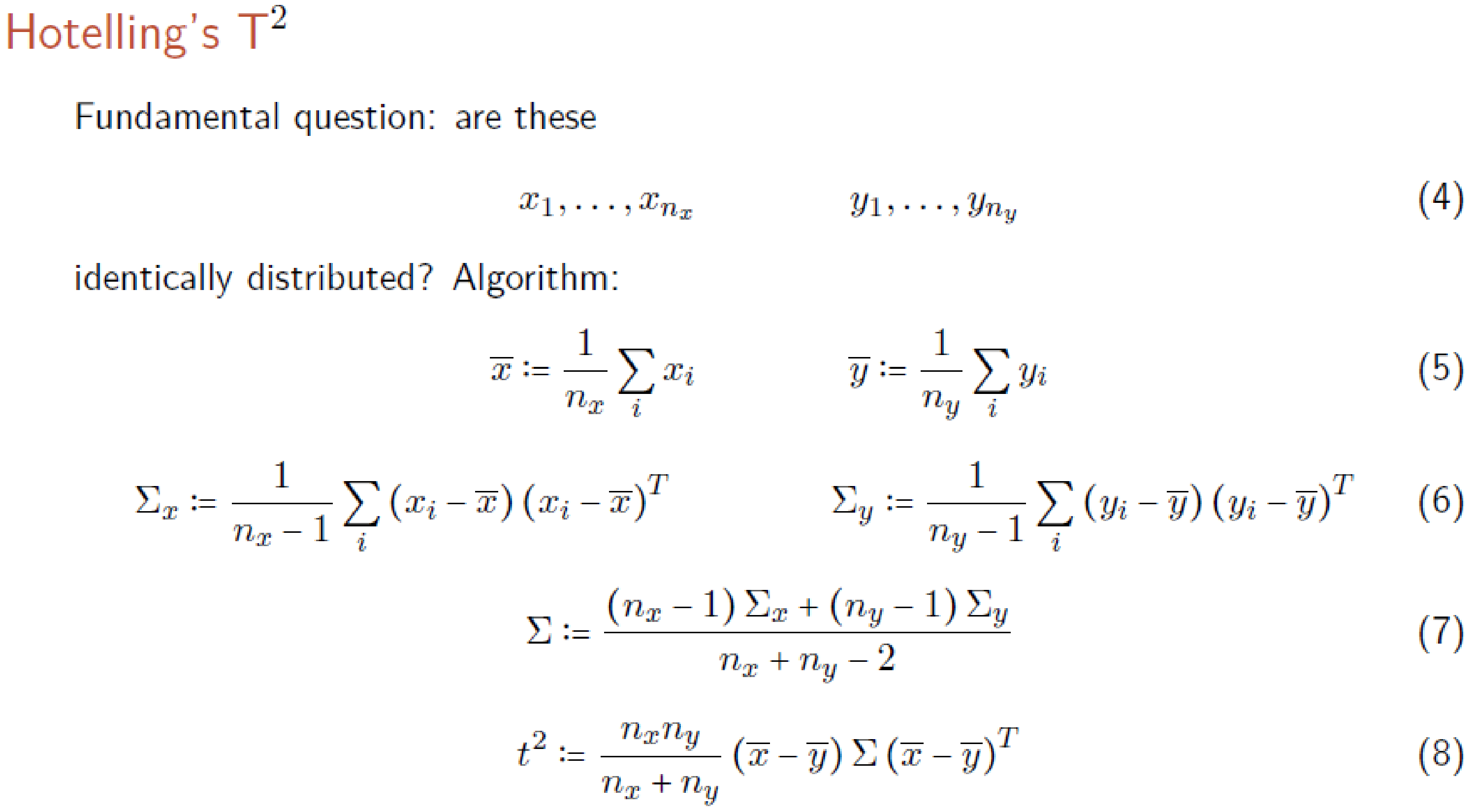


This sample has a large residual. Samples like these are not well described by the PCA model.

This sample has a high leverage. Such samples are influential to the PCA model.

Hotelling T^2 statistics

- A multivariate t-test
- Taken from Damiano's lecture:



Hotelling's T^2

Fundamental question: are these

$$x_1, \ldots, x_{n_x} \qquad y_1, \ldots, y_{n_y} \tag{4}$$

identically distributed? Algorithm:

$$\overline{x} := \frac{1}{n_x} \sum_i x_i \qquad \overline{y} := \frac{1}{n_y} \sum_i y_i \tag{5}$$

$$\Sigma_x := \frac{1}{n_x - 1} \sum_i (x_i - \overline{x}) (x_i - \overline{x})^T \qquad \Sigma_y := \frac{1}{n_y - 1} \sum_i (y_i - \overline{y}) (y_i - \overline{y})^T \tag{6}$$

$$\Sigma := \frac{(n_x - 1) \Sigma_x + (n_y - 1) \Sigma_y}{n_x + n_y - 2} \tag{7}$$

$$t^2 := \frac{n_x n_y}{n_x + n_y} (\overline{x} - \overline{y}) \Sigma (\overline{x} - \overline{y})^T \tag{8}$$

- The multivariate methods use the scores and not individual variables as input

Outliers in PCA (1/4)

- Leverage or Hotelling’s T^2 statistics: The distance from the mean of the model inside the model space

$$Hotelling's\,T_i^2 = \left(\sum_{a=1}^{A} \mathbf{t}_{ia}(\mathbf{T}_a^T\mathbf{T}_a)^{-1}\mathbf{t}_i^T\right)(I_{cal}-1)$$

The Hotelling’s T^2 statistic is approximately F-distributed as follows,

$$F_{A,I,\alpha} \sim T^2\frac{(I-A)}{A(I-1)}$$

- Leverage

$$h_i = 1/I_{cal} + \mathbf{t}_{ia}(\mathbf{T}_A^T\mathbf{T})_A^{-1}\mathbf{t}_{ia}^T$$

- The relationship between Leverage and Hotelling’s T^2

$$Hotelling's\,T^2 = (I_{cal}-1)(h-1/I_{cal})$$

- The critical limits Leverage and Hotelling’s T^2 are estimated from the scores from the model

Outliers in PCA (2/4)

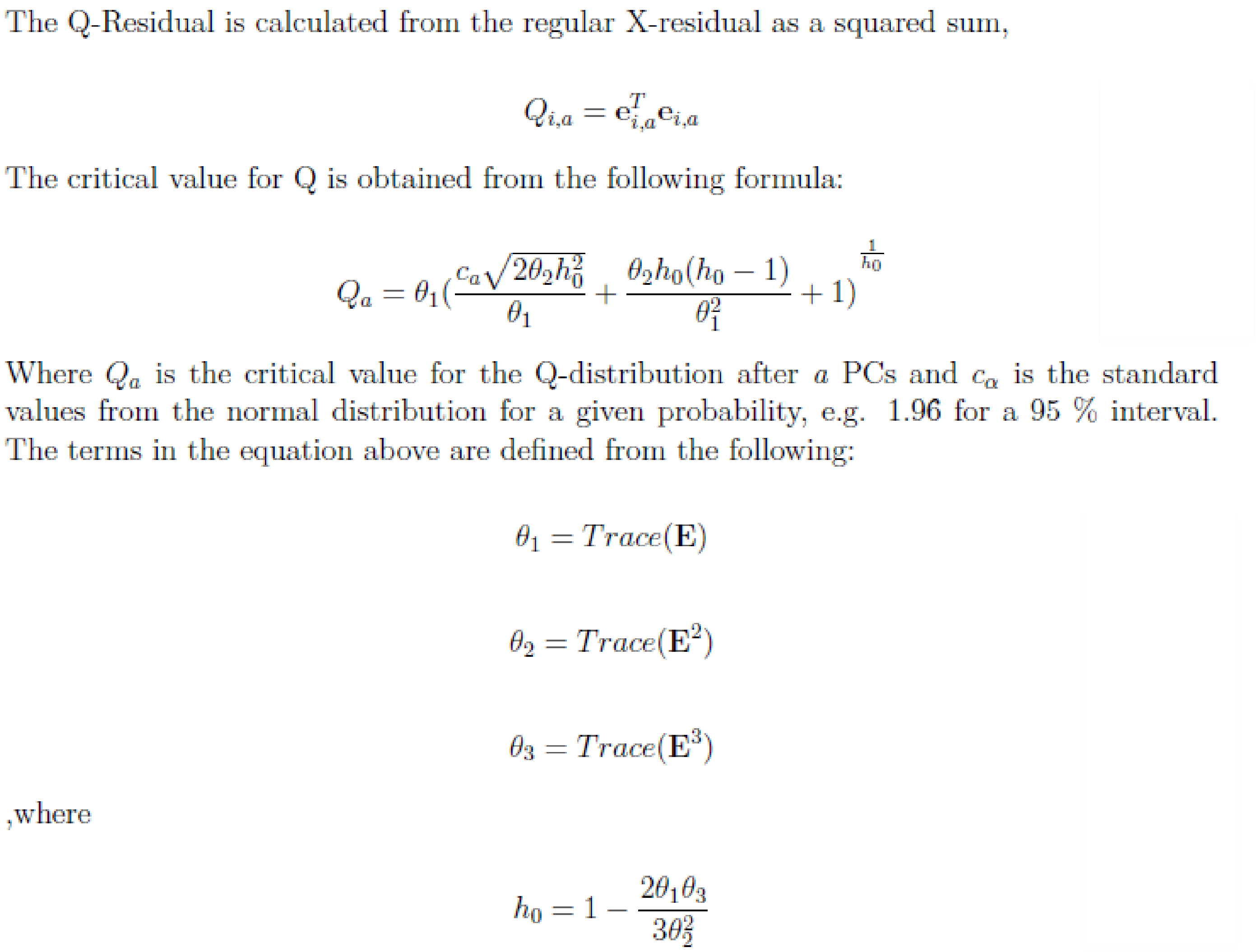
- The residuals can be represented in three ways:
  - As “raw” residuals: either a row or column in the **E** matrix
  - As sample residuals
  - As variable residuals

- The critical limits for sample residuals can be estimated from the F- or Q-distribution

- The contribution plots shows the individual variables’ contribution to the Hotelling’s T^2 statistic or the residuals

Outliers in PCA (3/4)

Q-residuals:



The Q-Residual is calculated from the regular X-residual as a squared sum,

$$Q_{i,a} = \mathbf{e}_{i,a}^T \mathbf{e}_{i,a}$$

The critical value for Q is obtained from the following formula:

$$Q_a = \theta_1\left(\frac{c_a\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2} + 1\right)^{\frac{1}{h_0}}$$

Where Q_a is the critical value for the Q-distribution after a PCs and c_α is the standard values from the normal distribution for a given probability, e.g. 1.96 for a 95 % interval. The terms in the equation above are defined from the following:

$$\theta_1 = Trace(\mathbf{E})$$

$$\theta_2 = Trace(\mathbf{E}^2)$$

$$\theta_3 = Trace(\mathbf{E}^3)$$

,where

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

Outliers in PCA (4/4)

F-residuals are closely connected to Analysis of Variance (ANOVA). The F-residual for an object *i* after *a* PCs:

$$F_{i,a} = Q_i/K$$

For validation, the expression is:

$$F_{i,a} = Q_i/(K - A)$$

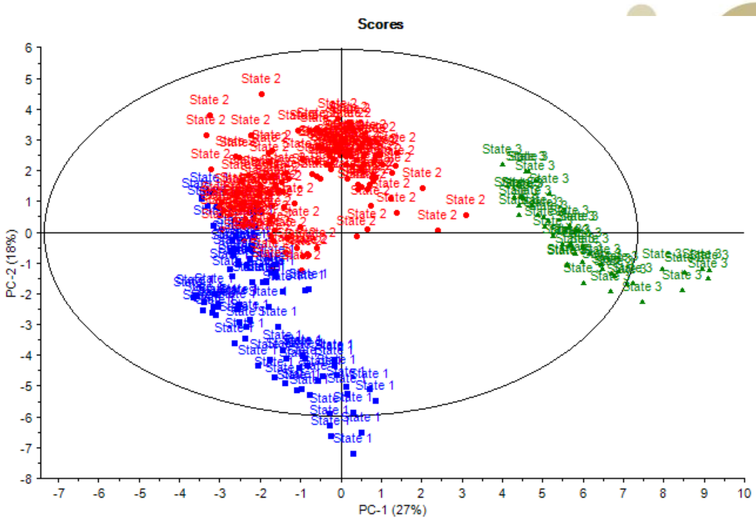
The term *(K-A)* is for correction of degrees of freedom when more PCs are included in the model (*K* = no. of variables, *A* = no. of PCs).

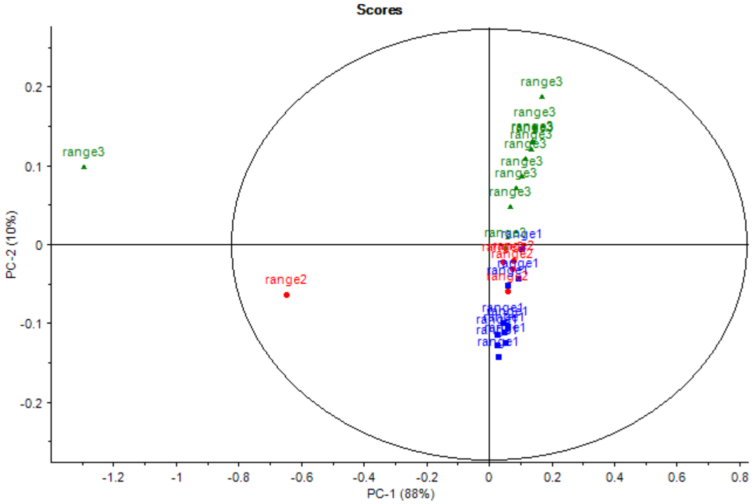
Q-residuals are often applied in Multivariate Statistical Process Control (MSPC) in situations with many samples and few variables.

For multichannel data, F-residuals are preferred

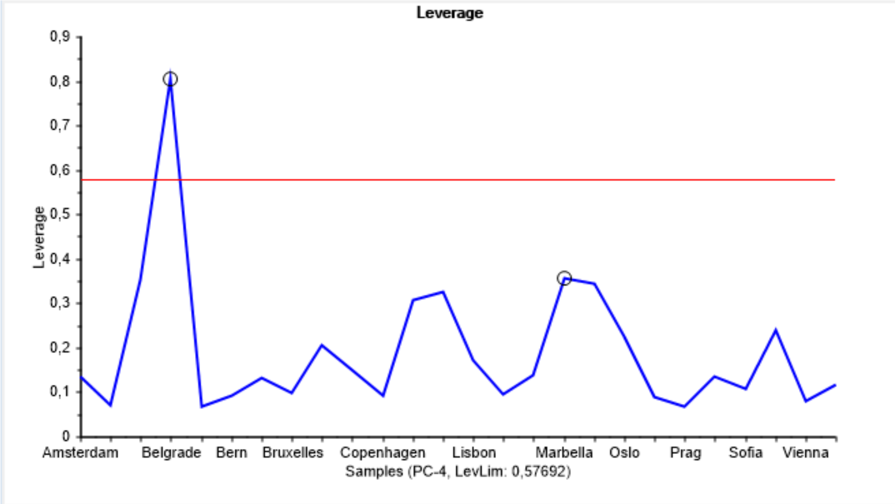
Hotelling T^2 ellipse

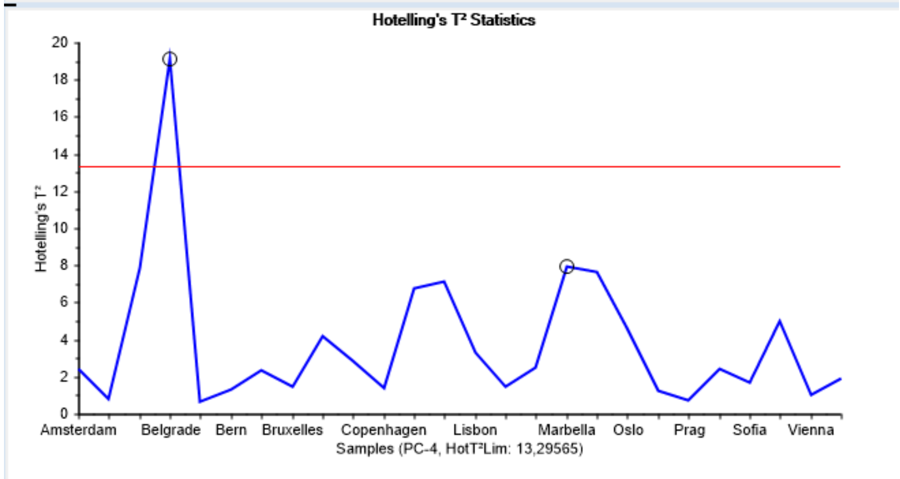
- The Hotelling T^2 ellipse is a 95% confidence limit
- The most influential samples tend to lie outside the ellipse
- *Useful for guidance, but do not blindly discard influential samples!*





Leverage and Hotelling's T²





- The leverage (h_i) of a sample describes how well the sample is described by the model, relative to the other samples
  - Leverage close to zero means the the sample is badly described by model
  - Leverage close to one means the sample may have high influence on the model.
- The Hotelling's T² statistic is a 'scaled' leverage and assumes a multivariate t-distribution

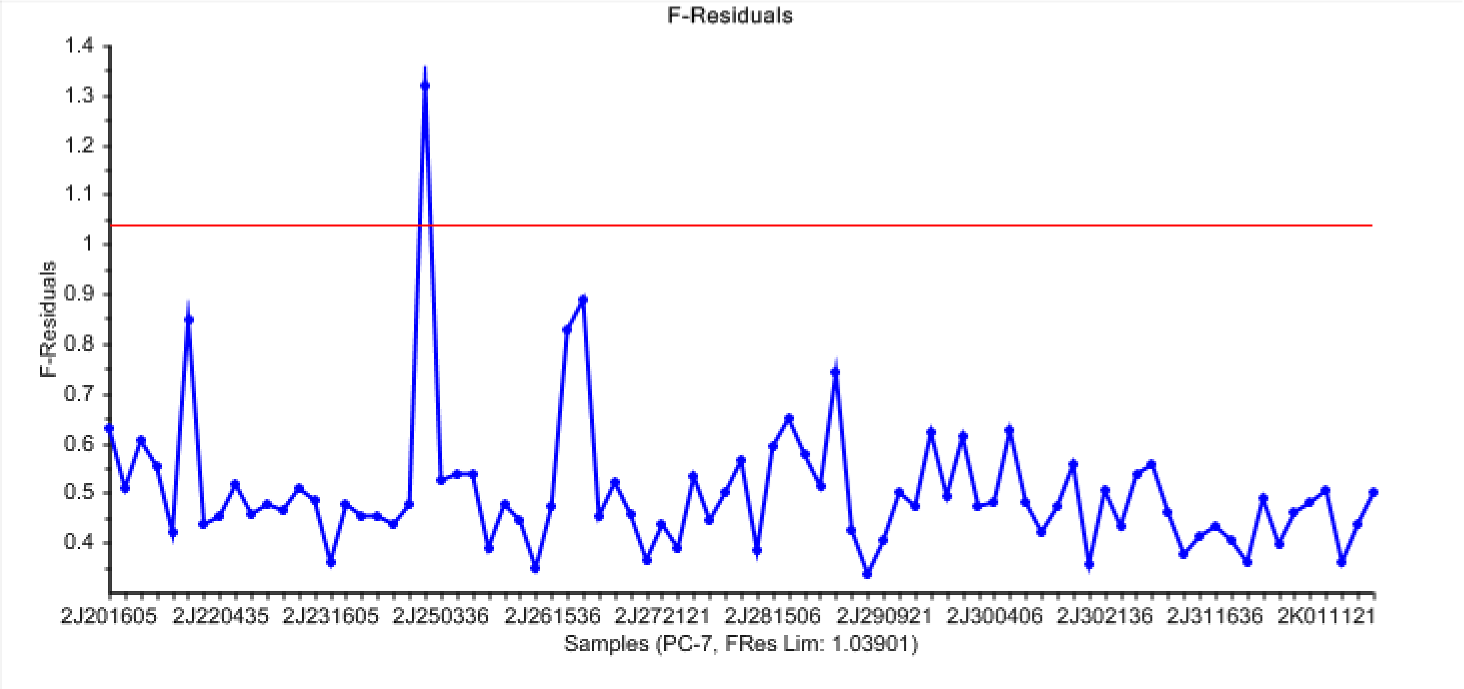
$$h_i = \frac{1}{n} + \sum_{a=1}^{A} \frac{t_{ia}^2}{t_a^T t}$$

The relationship between the two:

Hotelling's T² = $(n - 1)(h_i - 1/n)$

Residuals per sample

The critcal limits for the sample residuals can be estimated by Q- or F-residuals

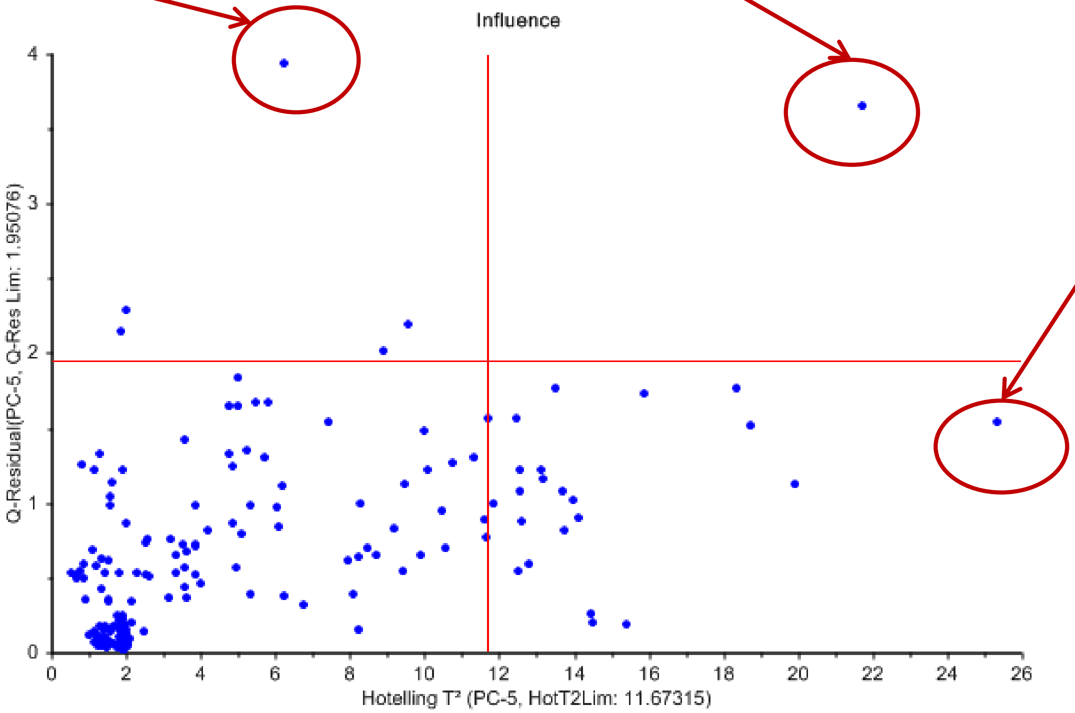


The Influence Plot – ”all in one”

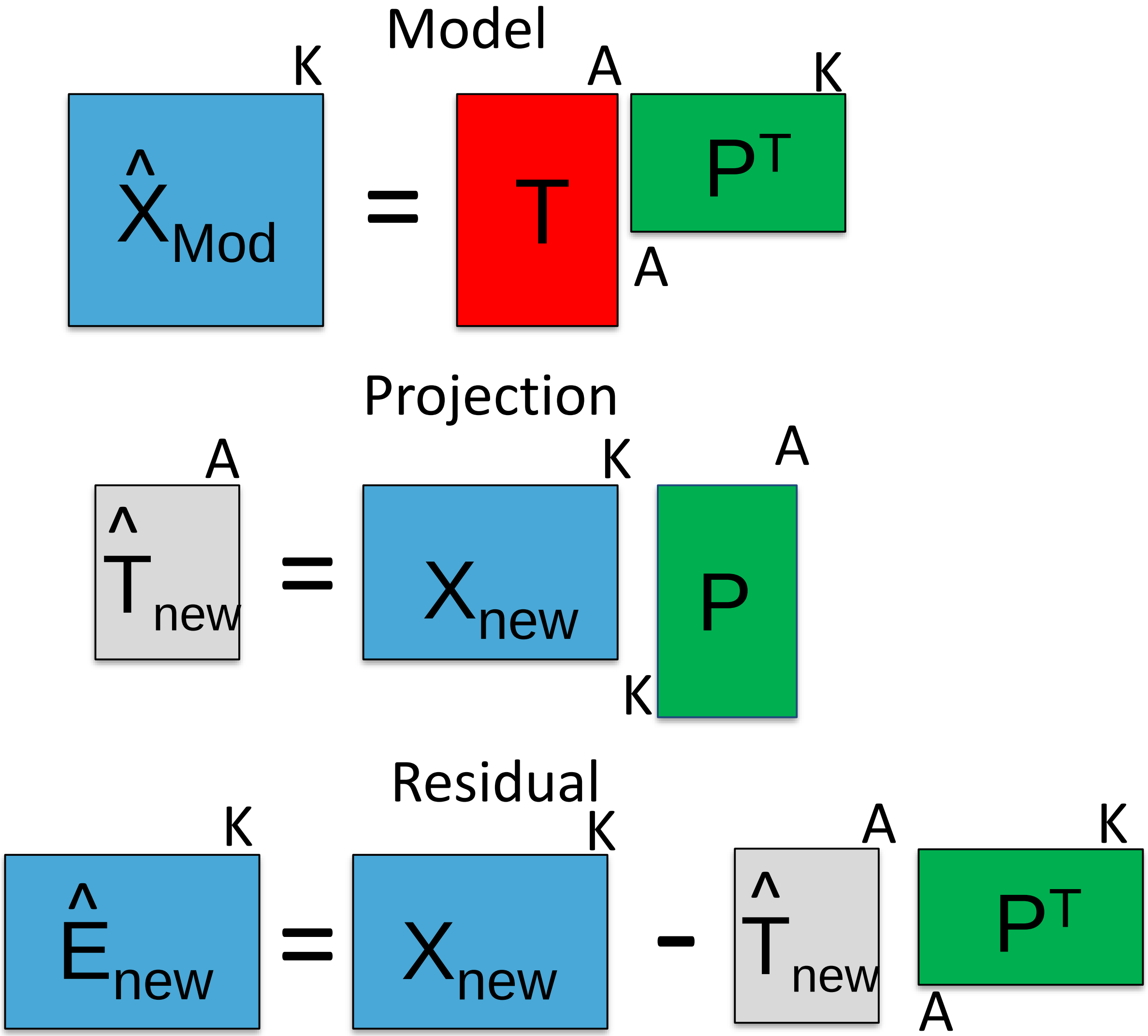
This sample has a large residual. Samples like this are not well described by the PCA model.

This sample has both a large residual and high leverage. Samples like this are most probably outliers.

This sample has a high leverage, and may overly influence the PCA model.



Projection of new samples



Model

$\hat{X}_{Mod} = T \; P^T$

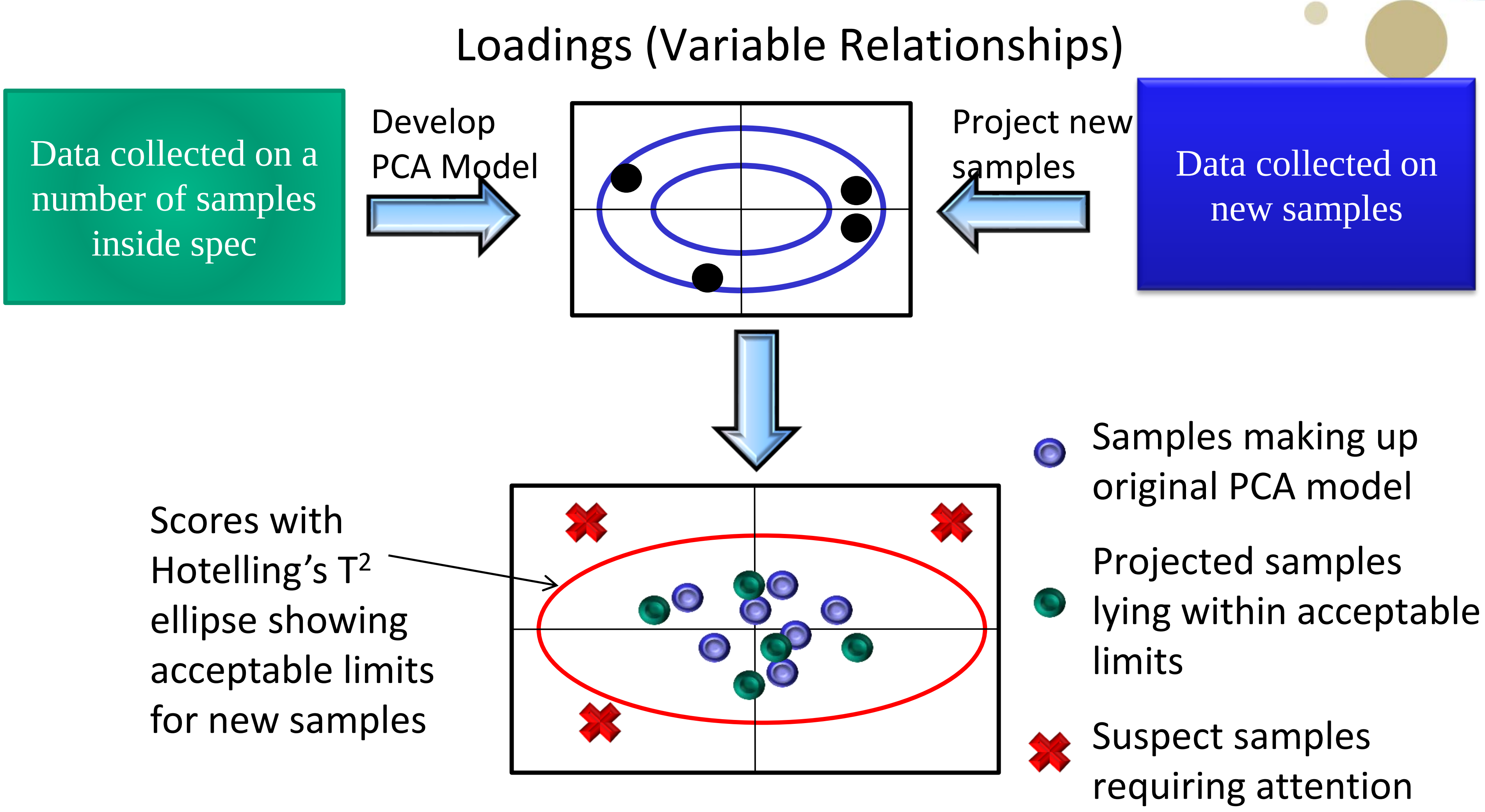
Projection

$\hat{T}_{new} = X_{new} \; P$

Residual

$\hat{E}_{new} = X_{new} - \hat{T}_{new} \; P^T$

Projection of new samples



Loadings (Variable Relationships)

Data collected on a number of samples inside spec

Develop PCA Model

Project new samples

Data collected on new samples

Scores with Hotelling’s T^2 ellipse showing acceptable limits for new samples

● Samples making up original PCA model

● Projected samples lying within acceptable limits

✖ Suspect samples requiring attention

Outliers in regression

X-outliers

- Can be spotted in PCA of X
- Can be spotted during regression

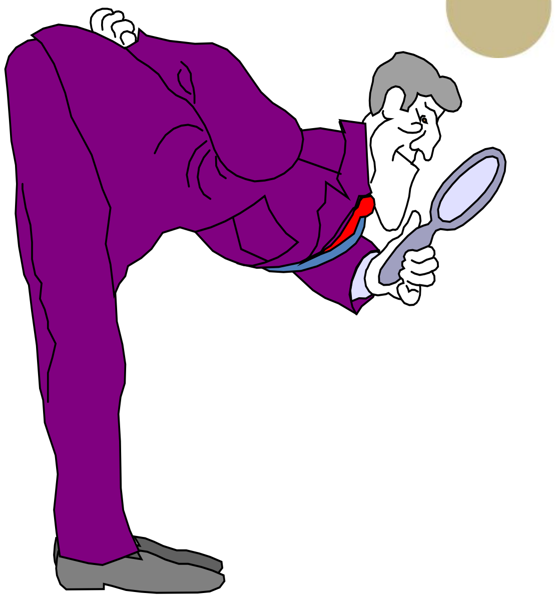
Y-outliers

- Start with PCA of Y if several responses
- Check range of Y using a histogram

X-Y Relation outliers (PLSR only)

- A sample may look OK with regards to X and Y separately, but its X-Y relationship (scores for X, scores for Y) may disagree with the global structure.

Outlier detection tools



Automatic distance measurements

- Residual variance
- Leverage
- Hotelling's T^2 statistic limit

Graphical tools

- Score plot
- X-Y relation outliers plot (T vs. U scores)
- Influence plot (Residual variance vs. Leverage/Hotelling's T^2)
- Normal probability plot of Y-residuals
- Y-Residuals vs. Predicted Y plot
- Predicted vs. Reference plot

Contribution statistics

When an outlier is detected it is of interest to identify which variables that contribute. For a sample $\boldsymbol{x}_i$ the contribution for each variable *k* after *A* components is (in pseudo code):

```
C(i,k) = 0;
for a = 1:A
    C(i,k) = C(i,k) + x(i,k).*p(k,a).*t(i,a)/Eigval(a);
end
```

The contribution to residuals can be shown as the values in $\boldsymbol{e}_i$ for the *K* variables or the square values of these

Contribution statistics in practice

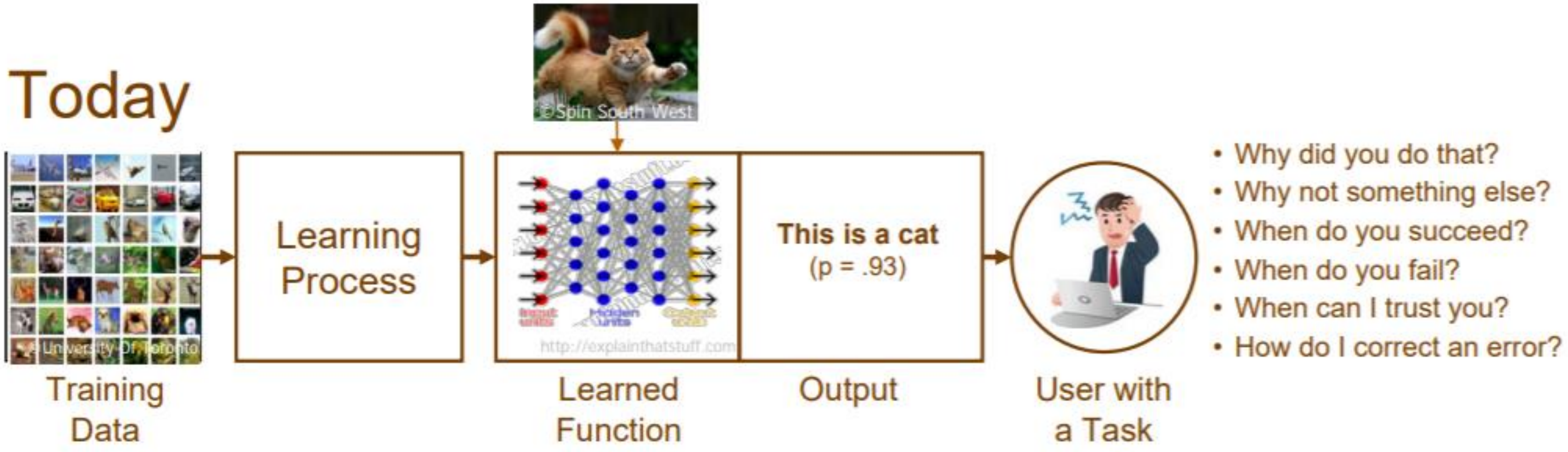
Procedure:

1. Visualize influence/score/Hotelling’s T^2/residual plots
2. Click on the sample outside critical limit
3. Show the contribution of individual variables (in model space and residual space)
4. Show time-line of selected variable

Process monitoring – detecting outliers in real-time applications

- Once a model has been established it may be used in real-time:
  - PCA: Is my process under control?
  - Classification: Is this the real Malaria medicine?
  - Regression: Quantitative prediction
  - Batch model: Is my fermentation proceeding as expected?
- Important aspects:
  - Edge analytics
  - Outlier detection
  - Interpretation
- The latent variable methods provide interpretation also during prediction (humans in the loop!)

Explainable AI (XAI)

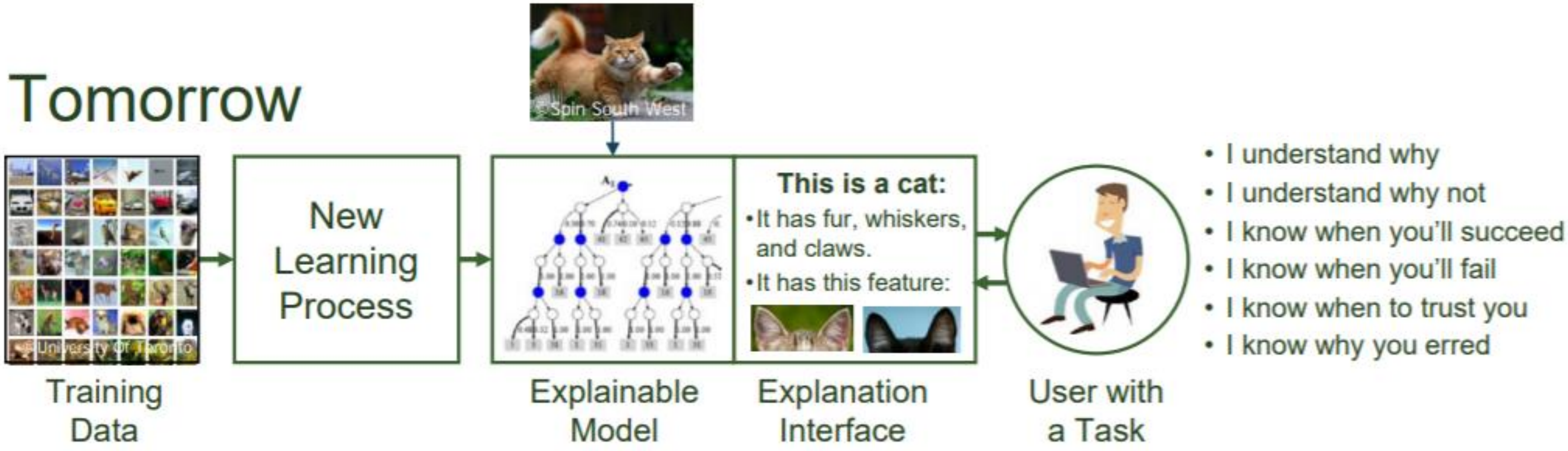


Today

Training Data → Learning Process → Learned Function → Output → User with a Task

This is a cat (p = .93)

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?



Tomorrow

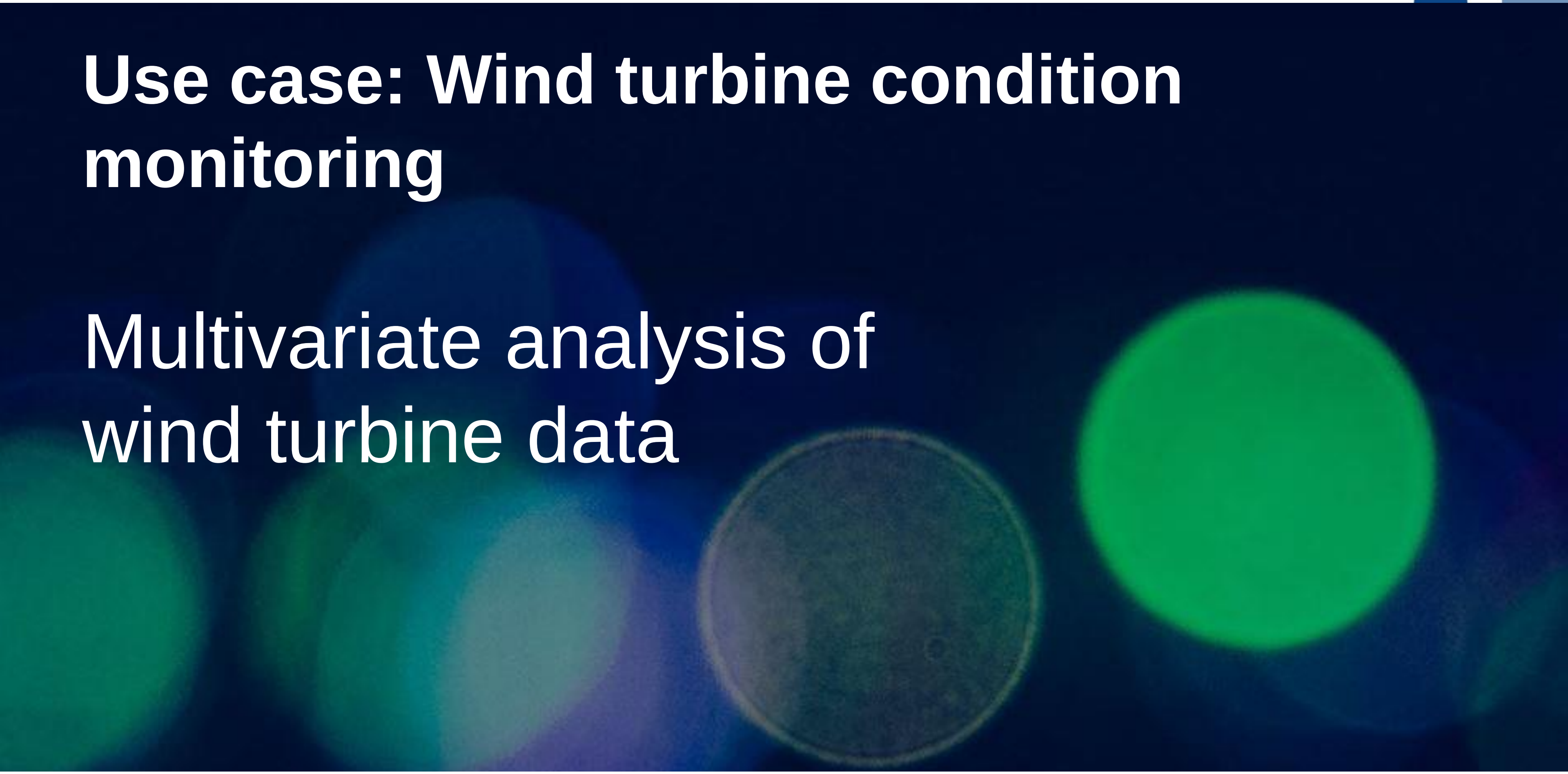
Training Data → New Learning Process → Explainable Model → Explanation Interface ⇄ User with a Task

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Use case: Wind turbine condition monitoring



Multivariate analysis of wind turbine data

Condition monitoring

The figure below illustrates the main parts of the wind turbine

A number of sensors are constantly monitoring the condition by measuring pressure, temperature etc.

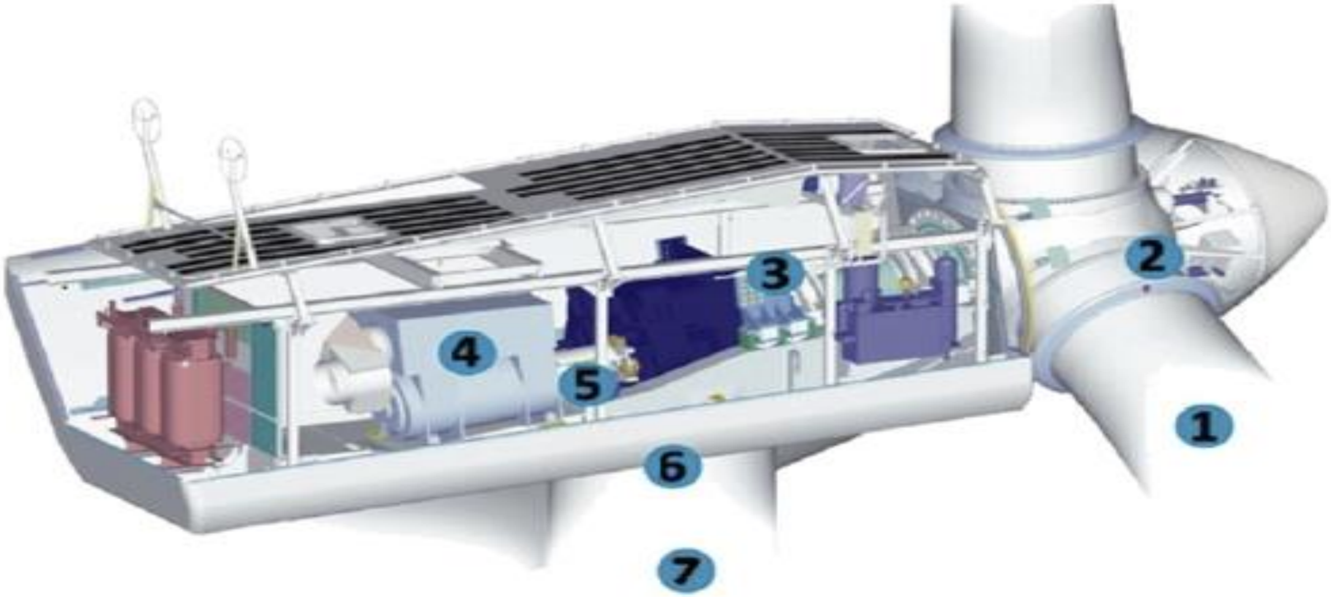


Fig. 3. Main parts of a turbine (Source: [11]) showing (1) blades, (2) rotor, (3) gearbox, (4) generator, (5) bearings, (6) yaw system and (7) tower.

Wind turbine monitoring – the multivariate advantage

- Many sensors in the wind turbine will give redundant information about the conditions.
- Thus, although with 200-300 individual sensors and derived variables the inherent dimensionality is much lower, i.e. there are correlations between sensors that are located on a specific part of the turbine and these can be "summarized" in terms of the underlying state
- Previous experience shows that the sensor signals will be grouped according to the variables mentioned above:
  - Condition of gear box, rotor,..
  - Sensors describing the yaw and pitch (active control of the turbine)
  - External conditions (temperature, wind etc.)
- This means that the turbine condition can be represented by a low (3-4) dimensional representation of the individual sensors.

Application example: Wind turbines

- Data:
  - 3093 observations
  - 275 variables
  - 7 turbines
- Establishing a model to be applied for real-time monitoring using a subset of the variables (but keeping all relevant information)
- Projecting the new observations with statistical limits and showing trend plots in the multivariate space. Interactive drill-down to find root cause in case of ”out-of-sweet-spot”

Analysis of wind turbine data

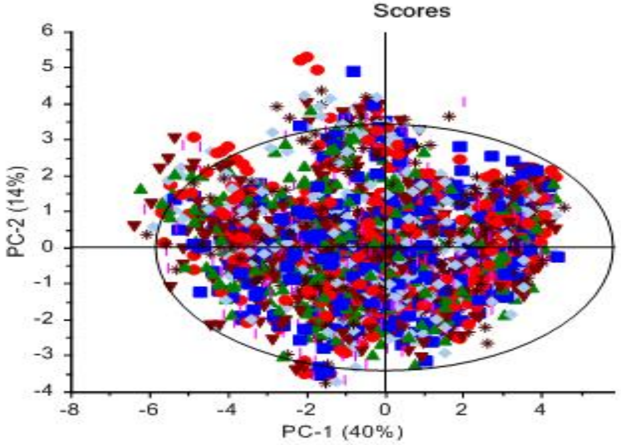
(14 variables only, for simplicity)

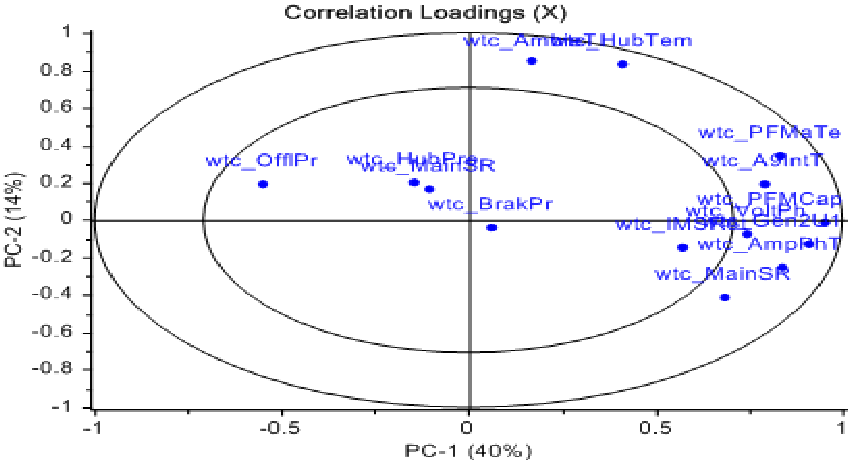
One overview gives information about all observations and variables

Map of observations

No differences between turbines

Variable relationships



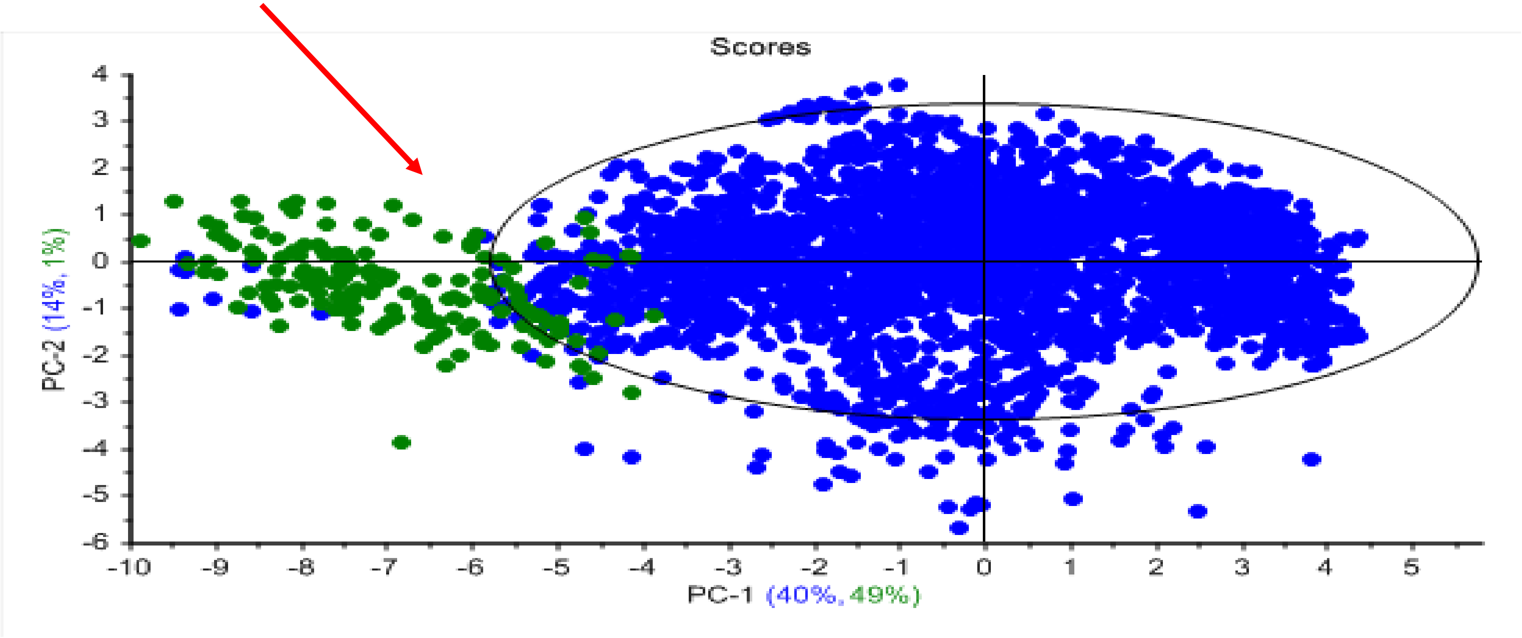


The model was validated across turbine

Projecting new observations off-line

Showing deviation from the ”sweet-spot”

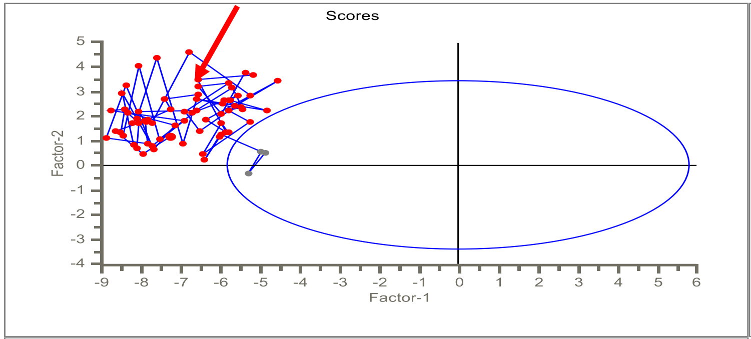
New observations in green

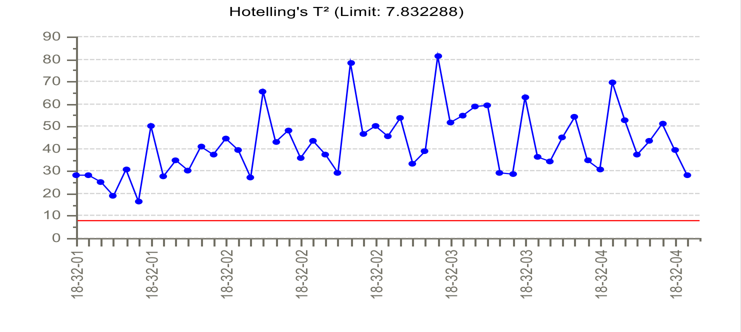


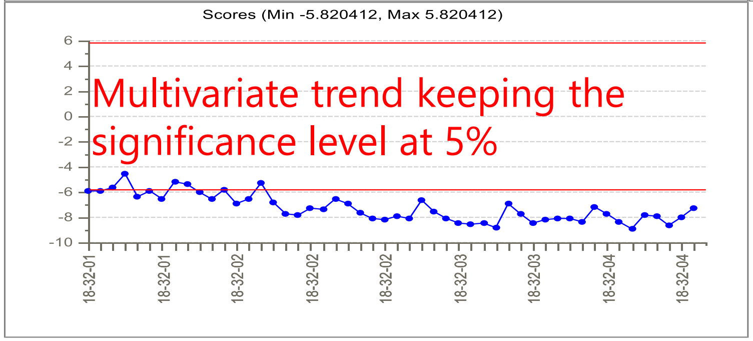
Real-time monitoring (simulated)

One overview gives information about new observations, i.e. possible failures

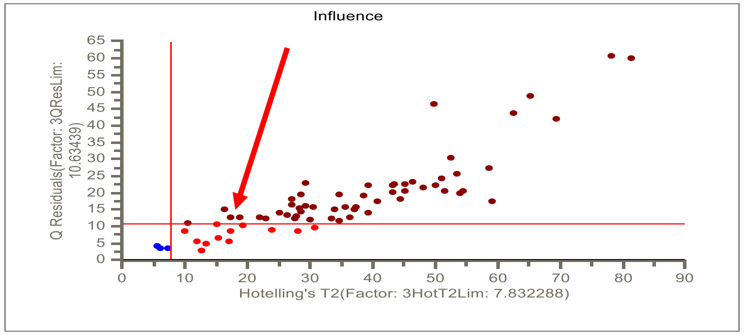
Observation is outside the normal condition ⇒ drill down to see why





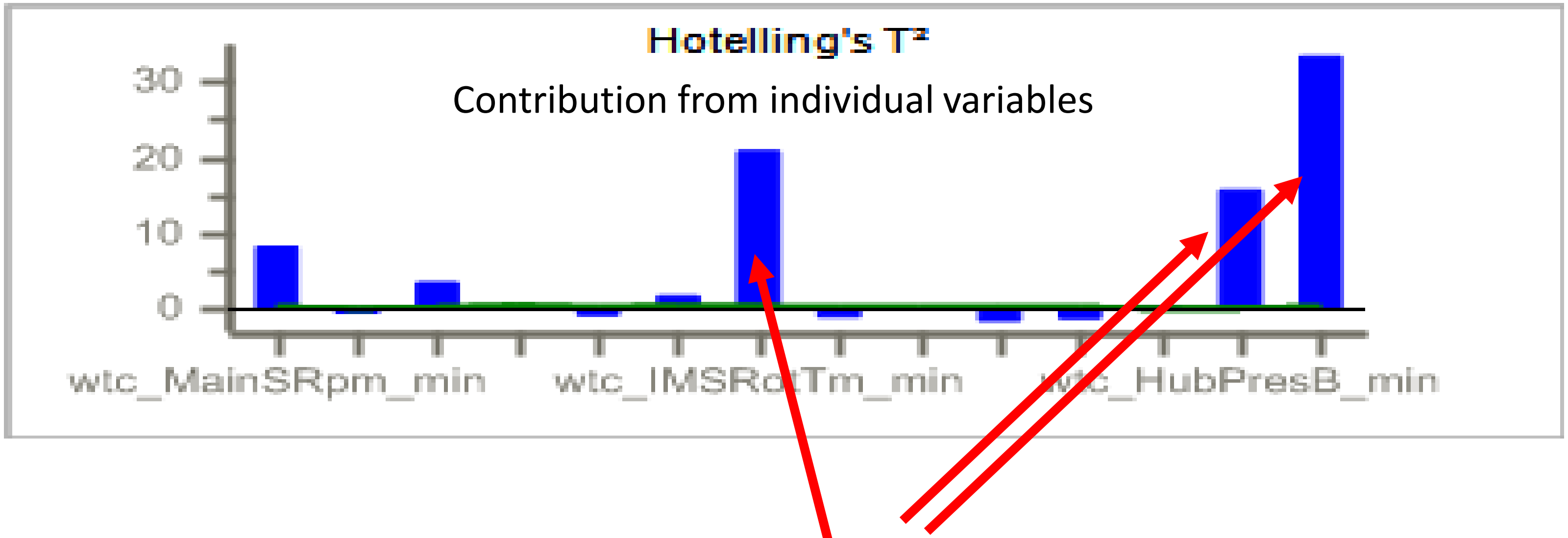


Multivariate trend keeping the significance level at 5%



Real-time monitoring

Contribution plot: Drill down to see root cause of change among many variables

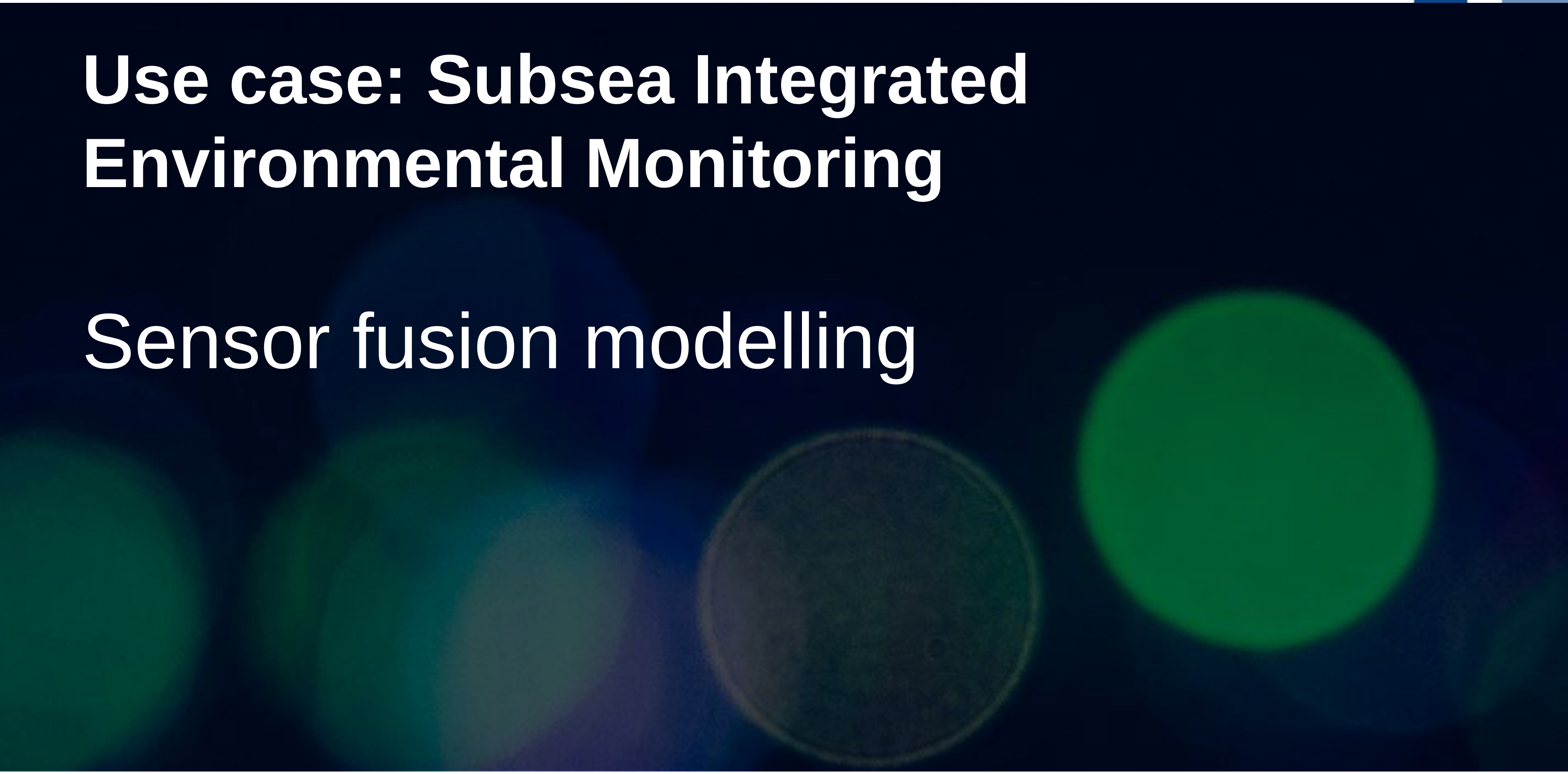


Contribution from individual variables

These variables have changed

Use case: Subsea Integrated Environmental Monitoring

Sensor fusion modelling

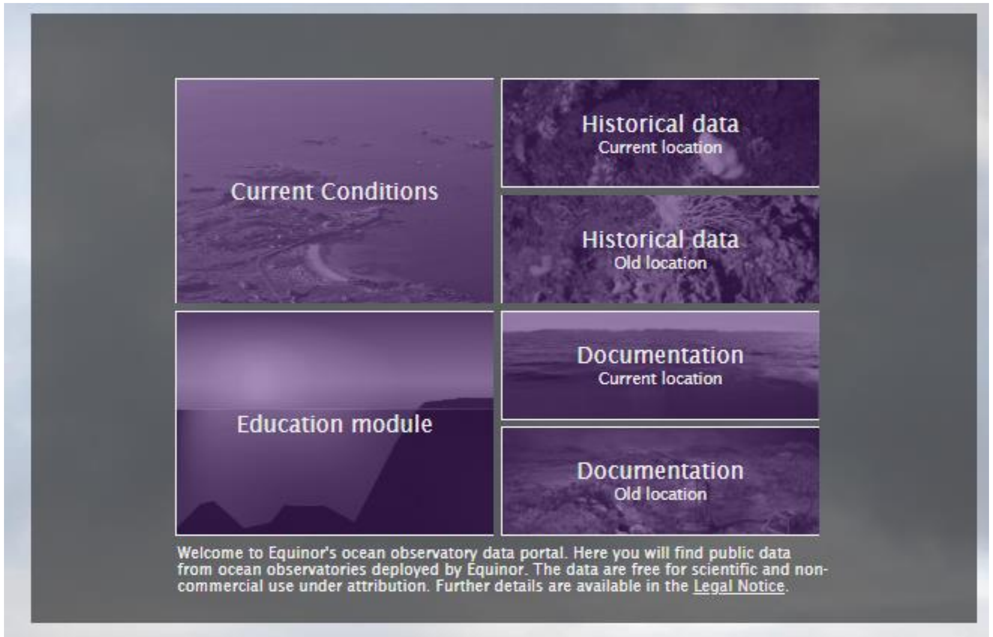


Environmental monitoring: The LoVe Ocean Observatory

- LoVe: Acronym for Lofoten-Vesterålen
- Lander with multiple sensors located at 258 m depth 20 km off the coast
- Cabled
- Data from the sensors (~100 individual variables) submitted online more or less continuously
- Real-time environmental monitoring

Ingvar Eide, Frank Westad, Automated multivariate analysis of multi-sensor data submitted online:
Real-time environmental monitoring. *PLoS One*, **13**, 1, 2018

The LoVe Ocean observatory



Current Conditions

Historical data
Current location

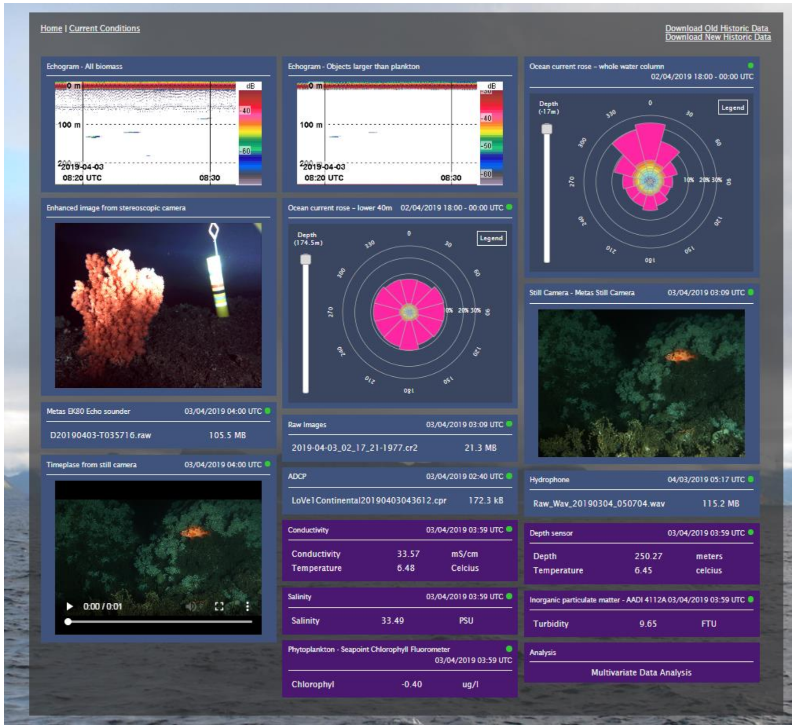
Historical data
Old location

Education module

Documentation
Current location

Documentation
Old location

Welcome to Equinor's ocean observatory data portal. Here you will find public data from ocean observatories deployed by Equinor. The data are free for scientific and non-commercial use under attribution. Further details are available in the Legal Notice.



Home | Current Conditions

Download Old Historic Data
Download New Historic Data

Echogram - All biomass

Echogram - Objects larger than plankton

Ocean current rose – whole water column
02/04/2019 18:00 - 00:00 UTC

Enhanced image from stereoscopic camera

Ocean current rose – lower 40m 02/04/2019 18:00 - 00:00 UTC

Still Camera - Metas Still Camera 03/04/2019 03:09 UTC

Metas EK80 Echo sounder 03/04/2019 04:00 UTC

D20190403-T035716.raw 105.5 MB

Raw Images 03/04/2019 03:09 UTC

2019-04-03_02_17_21-1977.cr2 21.3 MB

Timeplase from still camera 03/04/2019 04:00 UTC

ADCP 03/04/2019 02:40 UTC

LoVe1Continental20190403043612.cpr 172.3 kB

Hydrophone 04/03/2019 05:17 UTC

Raw_Wav_20190304_050704.wav 115.2 MB

Conductivity 03/04/2019 03:59 UTC

Conductivity	33.57	mS/cm
Temperature	6.48	Celcius

Depth sensor 03/04/2019 03:59 UTC

Depth	250.27	meters
Temperature	6.45	celcius

Salinity 03/04/2019 03:59 UTC

Salinity	33.49	PSU

Inorganic particulate matter - AADI 4112A 03/04/2019 03:59 UTC

Turbidity	9.65	FTU

Phytoplankton - Seapoint Chlorophyll Fluorometer
03/04/2019 03:59 UTC

Chlorophyl	-0.40	ug/l

Analysis

Multivariate Data Analysis

https://love.statoil.com/

Online sensors at LoVe

- A total of ~100 individual variables are collected every 5th or ten minutes:
  Chlorophyll (2 sensors), conductivity, depth, temperature (3 sensors), turbidity and Total Suspended Matter, salinity, biomass at three different depths, and current speed
- Current speed is measured in two directions (N and E) using two sensors covering different depths, however, with overlap:
  - Aquadopp 3-21 m, 2 m resolution
  - Continental 6-146 m, 5 m resolution
- RGB Camera

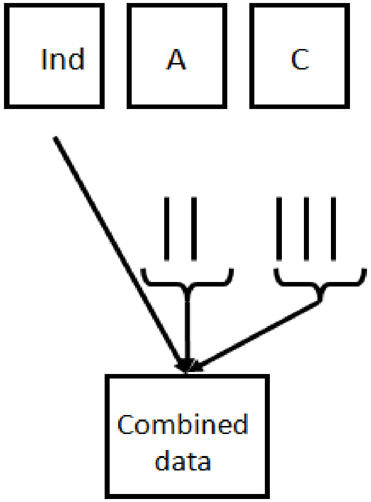
The modelling approach

Multiblock model

Variable blocks

Scores from blocks A &C

Perform PCA



Ind

A

C

Combined data

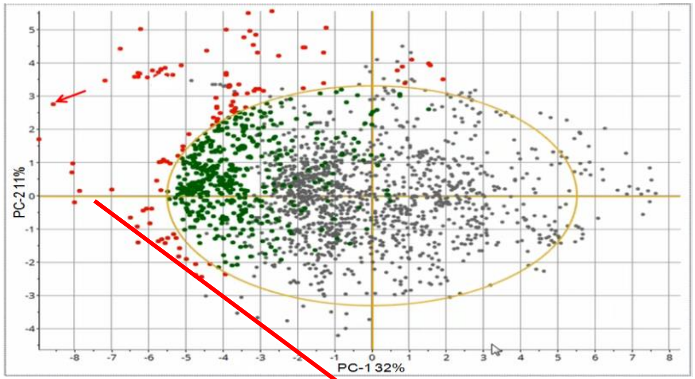
Procedure:

1. PCA on Aquadopp and Continental

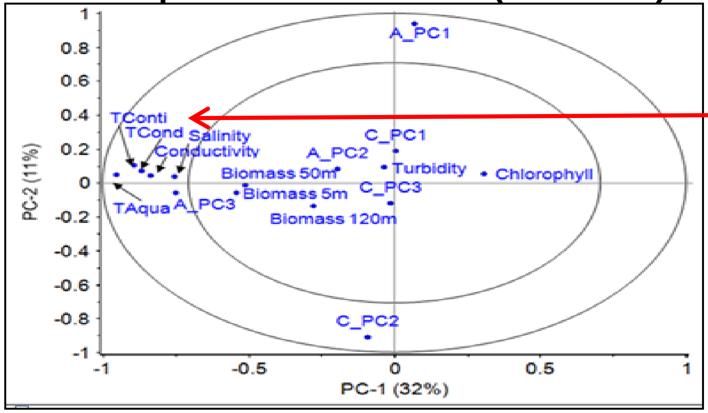
2. Combine individual sensors and scores in the final model

NB! Validation must be done for all steps!

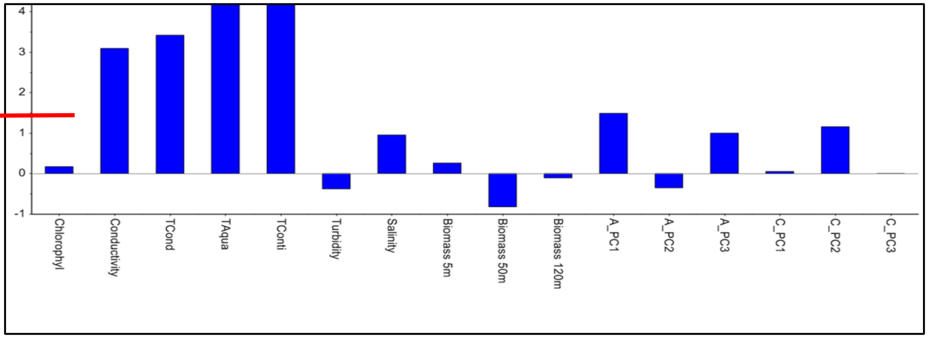
The LoVe case – on-line monitoring



Map of variables (model)



Drill down



Higher temperature in August

Regression models and the relationship to ARMAX models

- Assume an ARMAX model with three main terms:

AR MA X

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{b} \eta_i d_{t-i}.$$

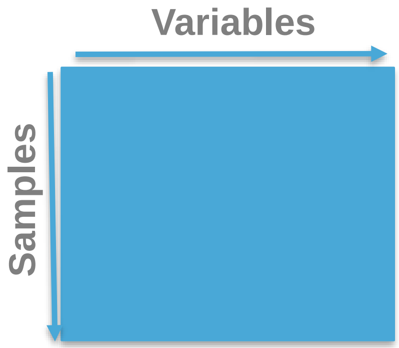
- In most multivariate models only the exogenous variables (X) are used, i.e. all variables collected in the process for many periods of time to reflect all known and unknown sources of variation
- However, one may also include AR terms, MA and lagged X-variables to reflect the dynamics of the time series

Batch Processes

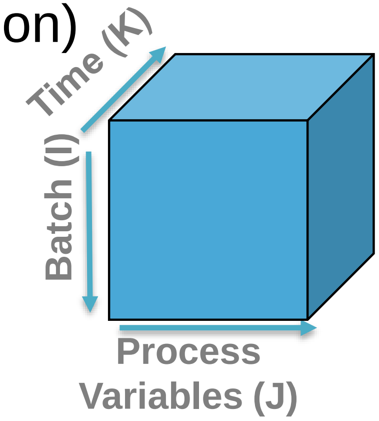
- Batch processes are common in chemical, pharma and biopharma industries
- Objectives:
  - How can I analyse the batch data with the goal of process optimisation (yield, duration)?
  - Can I find the reason why product quality for some batches lies outside the specifications?
  - Are there any effects from raw materials/season/operator/equipment?
- Multivariate batch modelling and monitoring can be applied for quality control and event detection

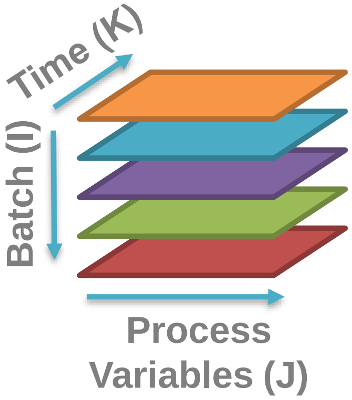
Multivariate Analysis (MVA) & Batch Data

- Traditional MVA- Two dimensional matrices (no time dimension)



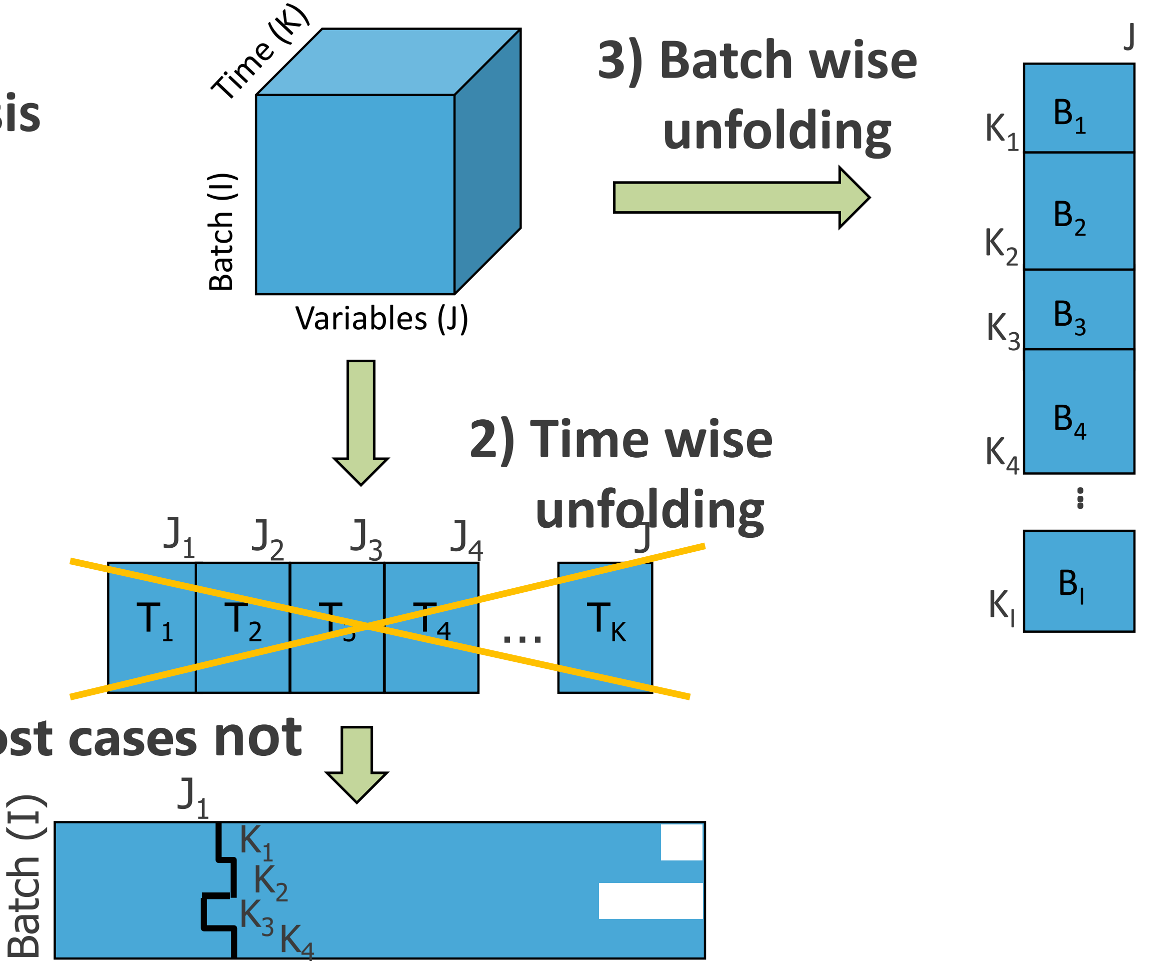
- Batch MVA- Three dimensional data (time as a third dimension)





Batch Data: Unfolding Strategies

1) Direct 3-way analysis



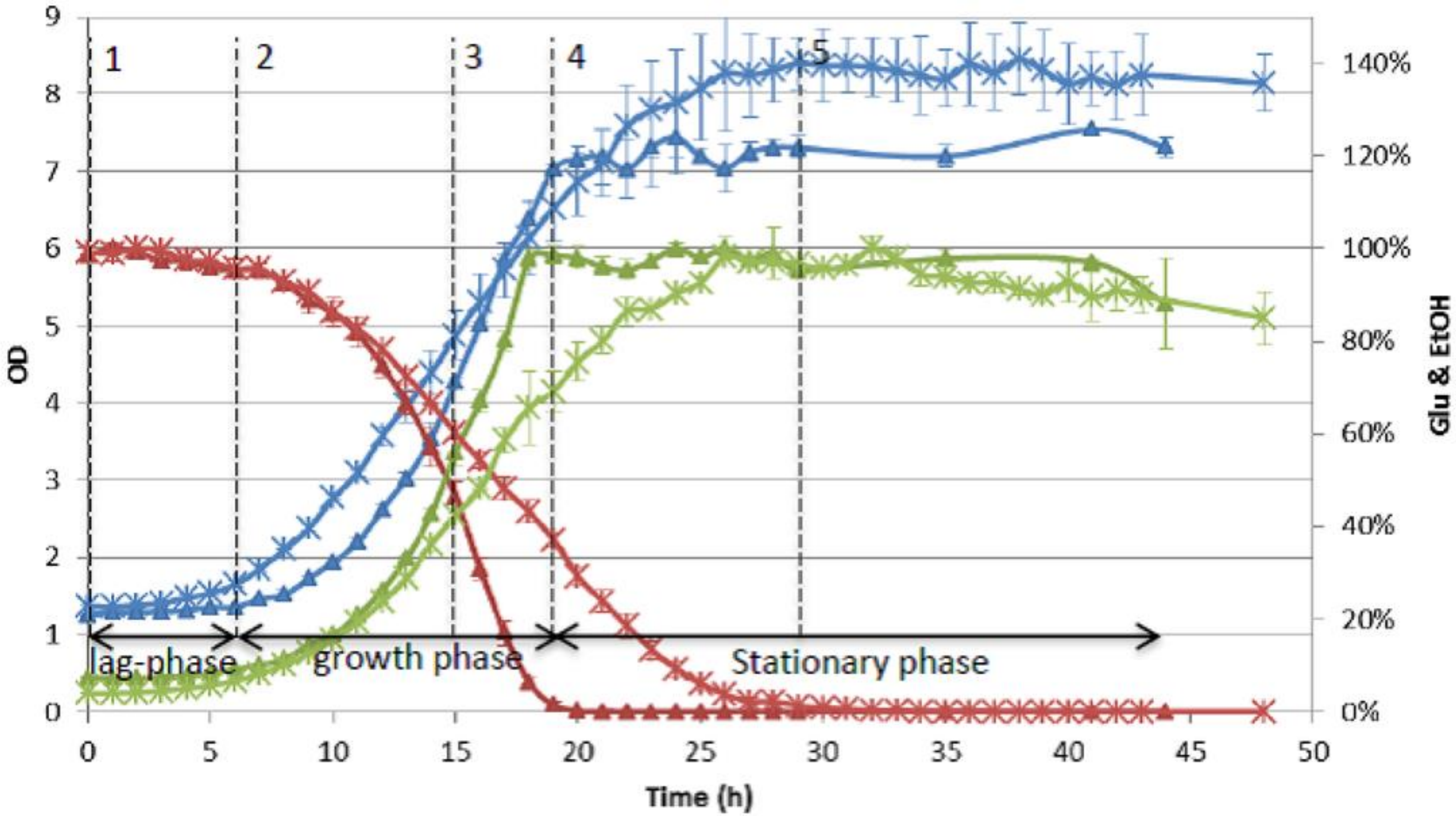
3) Batch wise unfolding

2) Time wise unfolding

K is in most cases not constant!

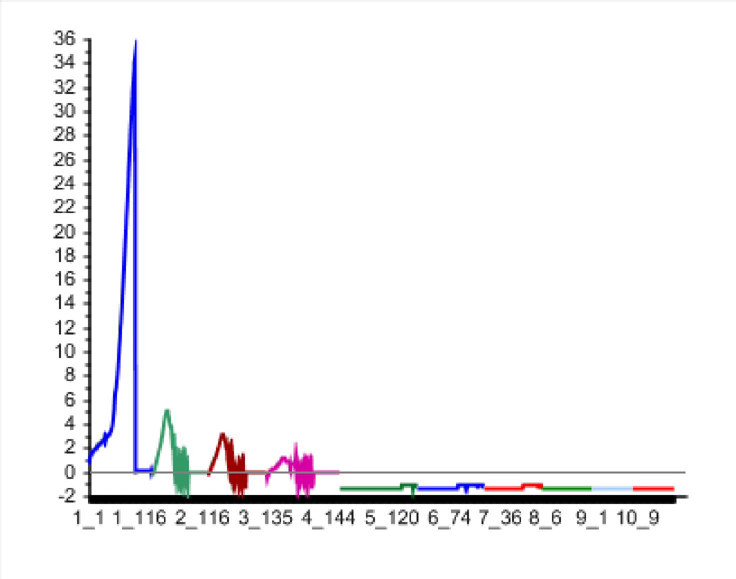
Example: Fermentation –

Two batches and three variables

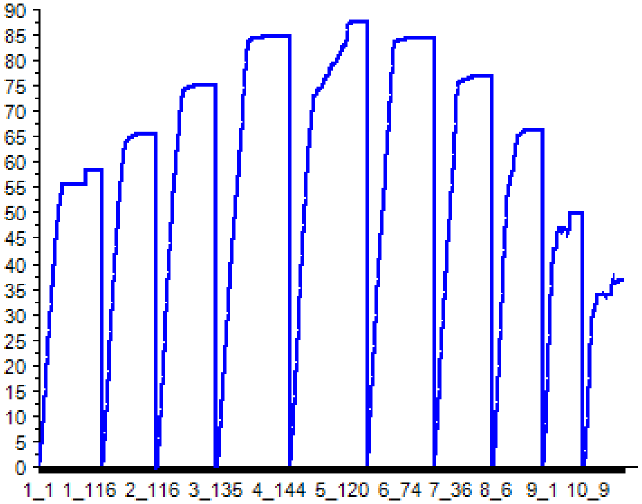


Batch Analysis Challenge: Controlled variables

Different set points across batches



Partially constant

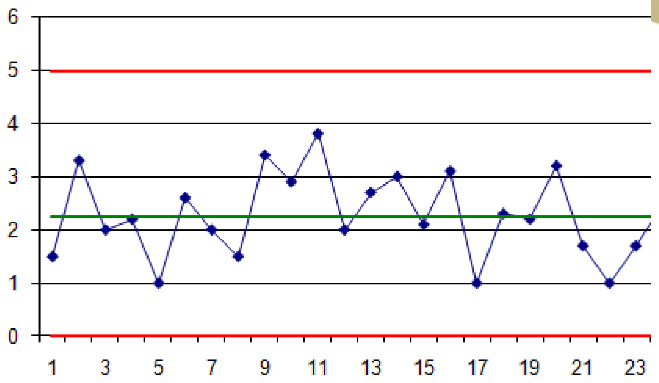


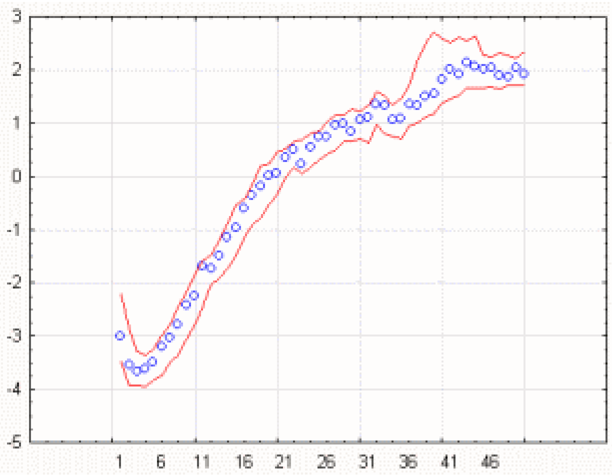
Batch modelling & monitoring

- Data for a number of batches are collected
- Step 1: Analyze all data; Explorative phase
  - Interpret variable structure; remove outliers
  - Make model on golden batches
  - Store model for on-line monitoring
- Step 2: On-line monitoring
  - Follow batch over time
  - Detect out-of-spec situations
  - Find cause (e.g. residuals and contribution plots)
  - Take action (process control/feedback system required)

Control charts –(M)SPC

- Univariate plots (with confidence limits) of
  - Process parameters
  - Multivariate model statistics or parameters
  - Multivariate predictions
- Batches are non-steady-state, i.e. both target and confidence limit change with time
- Because sample number is plotted on the x-axis, **unequal batch length and starting point** have to be taken into account





Synchronization and time warping

- Synchronization helps in defining start and end for various phases of the process
- Time warping helps in making all batches have the same length
- Taken from some recent publications:
  - “This paper is devoted to demonstrate that synchronization is a critical and necessary preliminary step to bilinear batch process Modelling, no matter whether batch trajectories have equal length or not”
  - “It is necessary to align the time differences for both the measured trajectories and the batch end-product quality in order to implement statistical process monitoring and control schemes”.
- But is this really necessary?

The Solution: Model the process in relative time

- Removes dependence of the process time axis
- Allows visualisation of how the process evolves independently of sampling rate
- Enables plot of individual variables in relative time
- Some additional comments:
  - Do not include variables that are kept constant or are changed in steps such as stirring, dosage etc. Keep these in a separate model.
  - Sensors such as spectrometers describe the chemical/biological state directly or indirectly

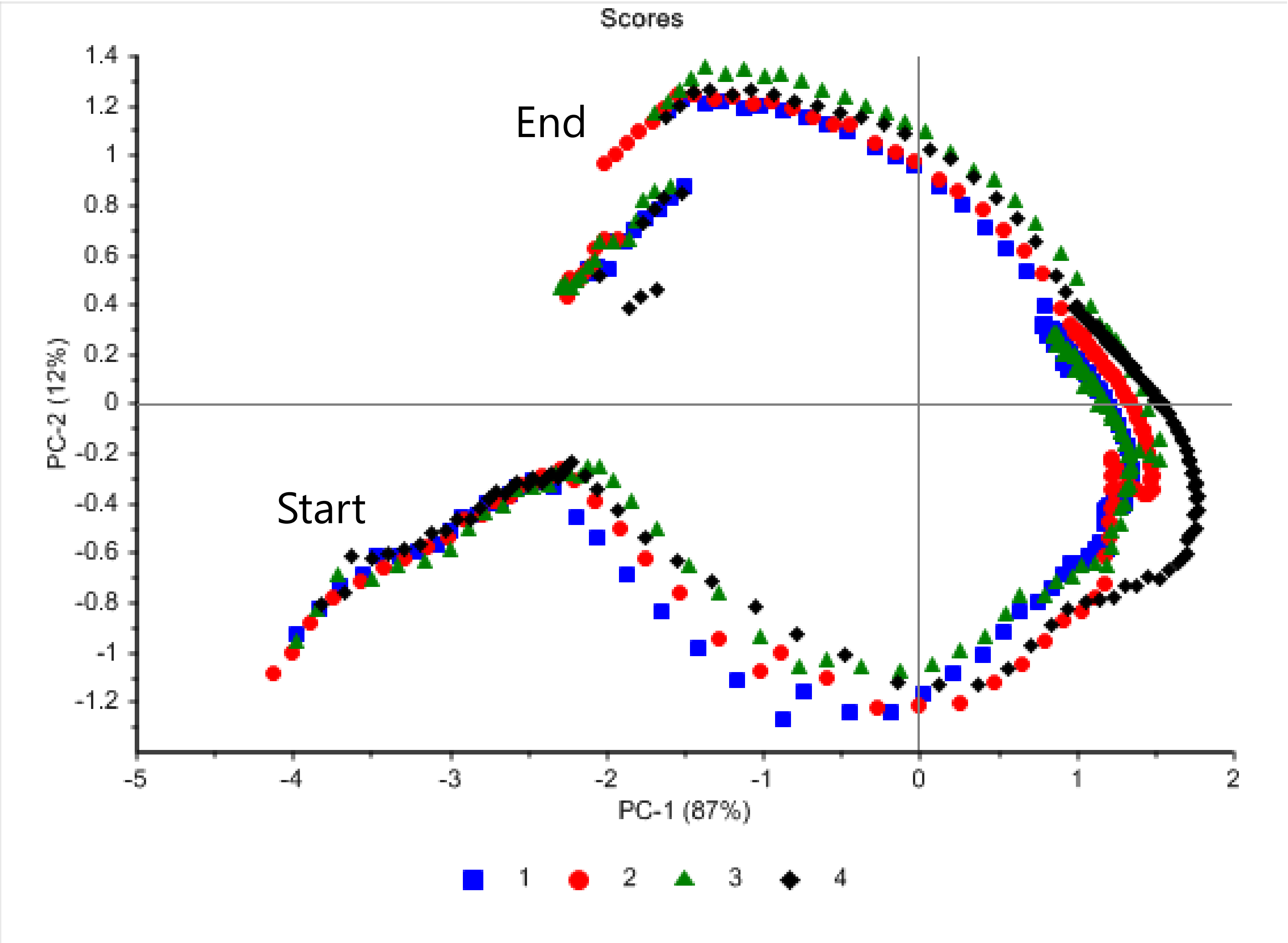
Monitoring a new batch

- This approach has a particular advantage in the monitoring phase:
  - Independent of sampling rate
  - Displays batch progression in relative time
  - Able to model non-linear behavior
- Details:
  - Monitor batch over time in score space
  - Detect out-of-spec situations
    - Trajectory model distance
    - Distance to model
    - Contribution plot
  - Take action (process control/feedback system required)

Example:

- Chemical reaction
- 4 Historical batches – PCA with validation across batch
- 3 variables, Temperature x 2, Pressure
- Projecting a new batch

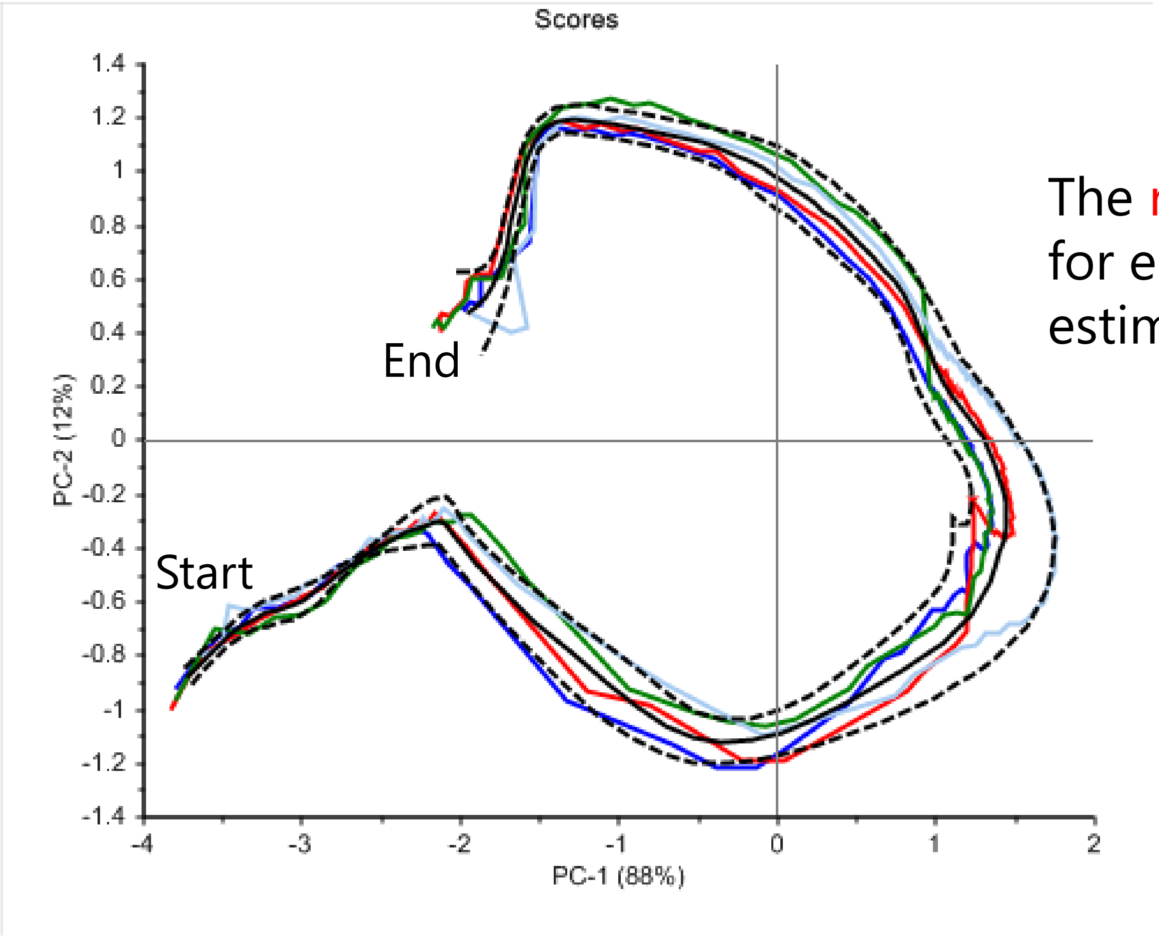
Scores from PCA for four historical batches



Chemical reaction

Average trajectory and 95% confidence interval

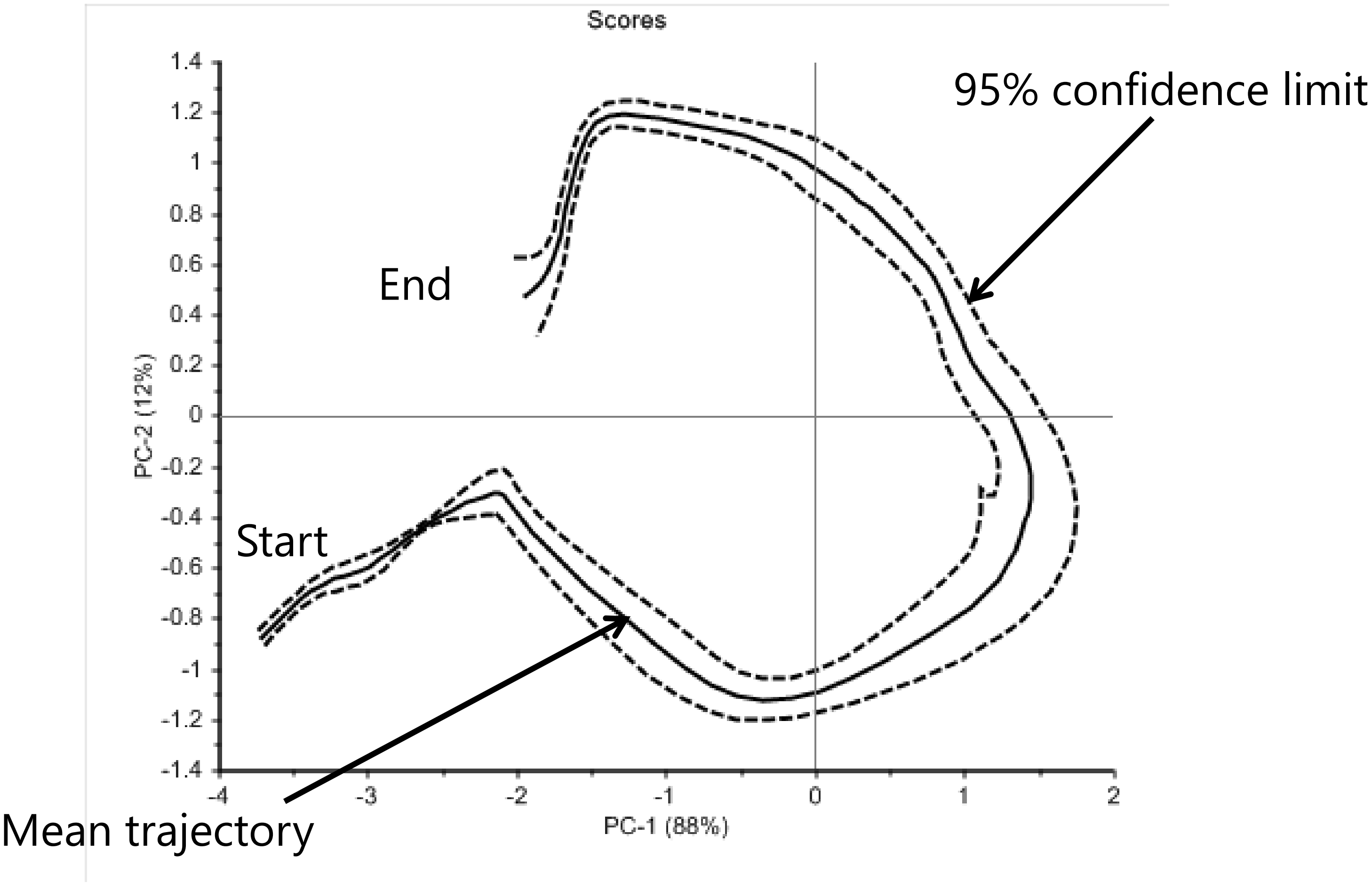
With historical batches



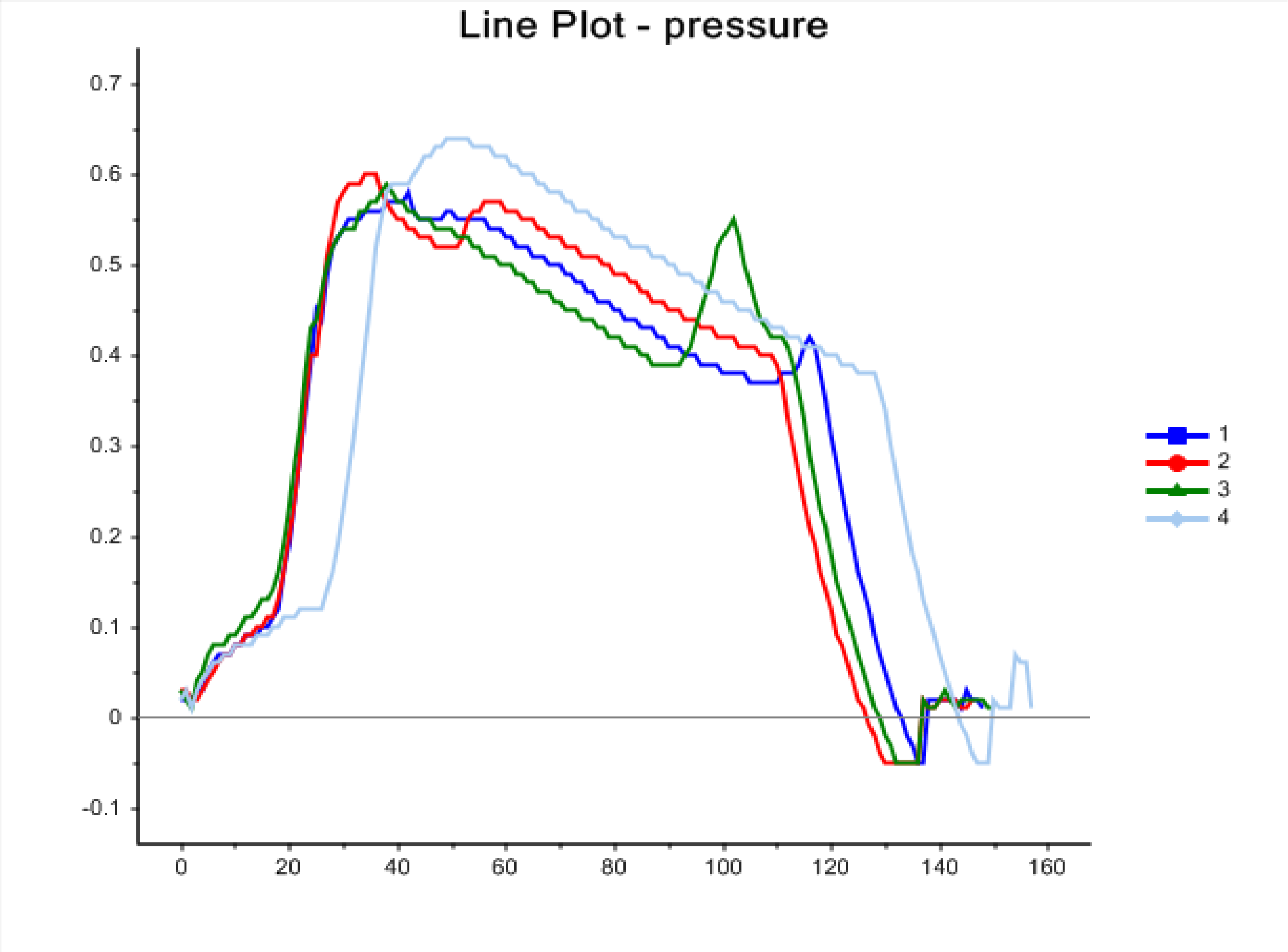
The relative time for each sample is estimated

Chemical reaction

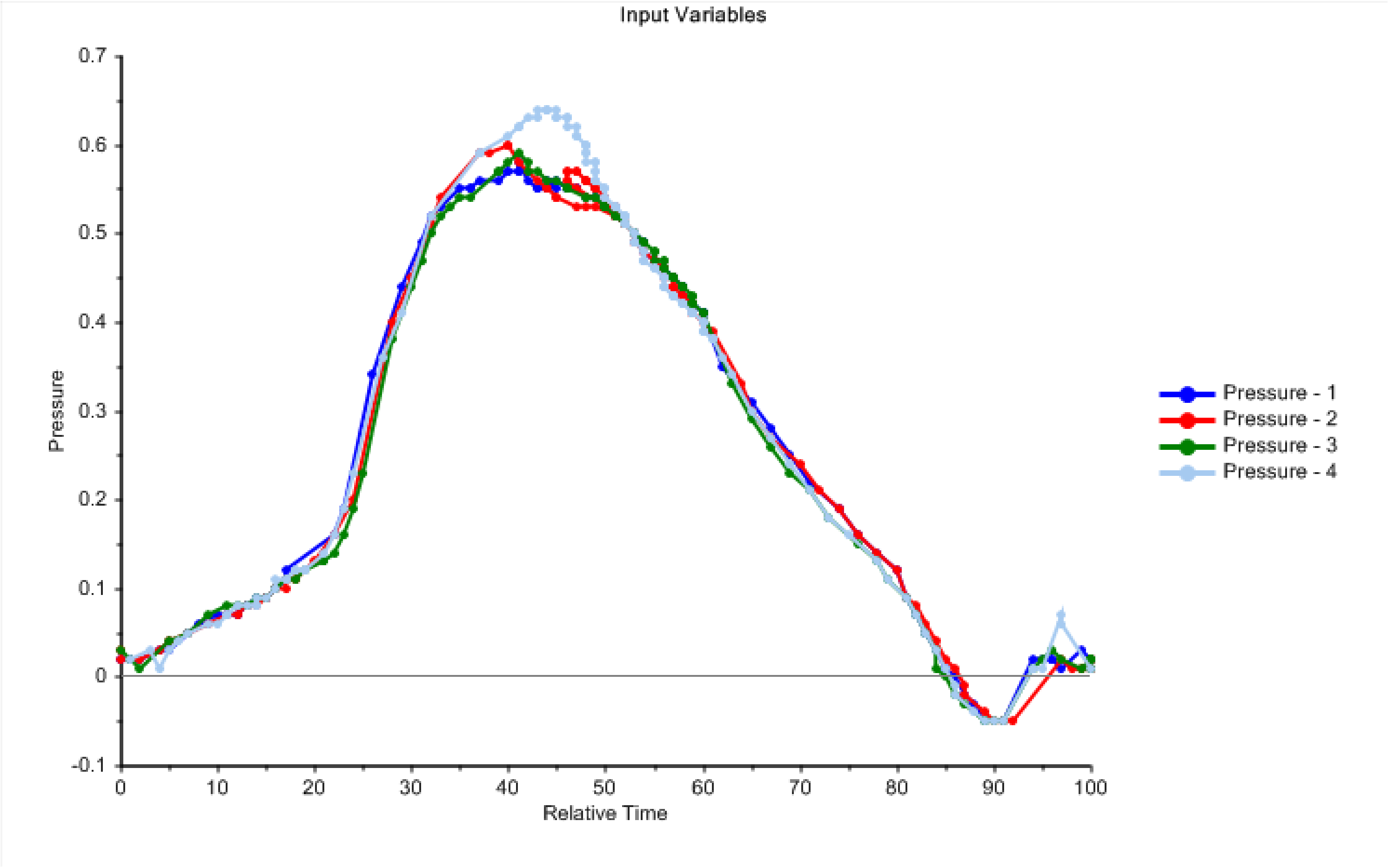
Average trajectory and 95% confidence interval



Pressure showed as sample number



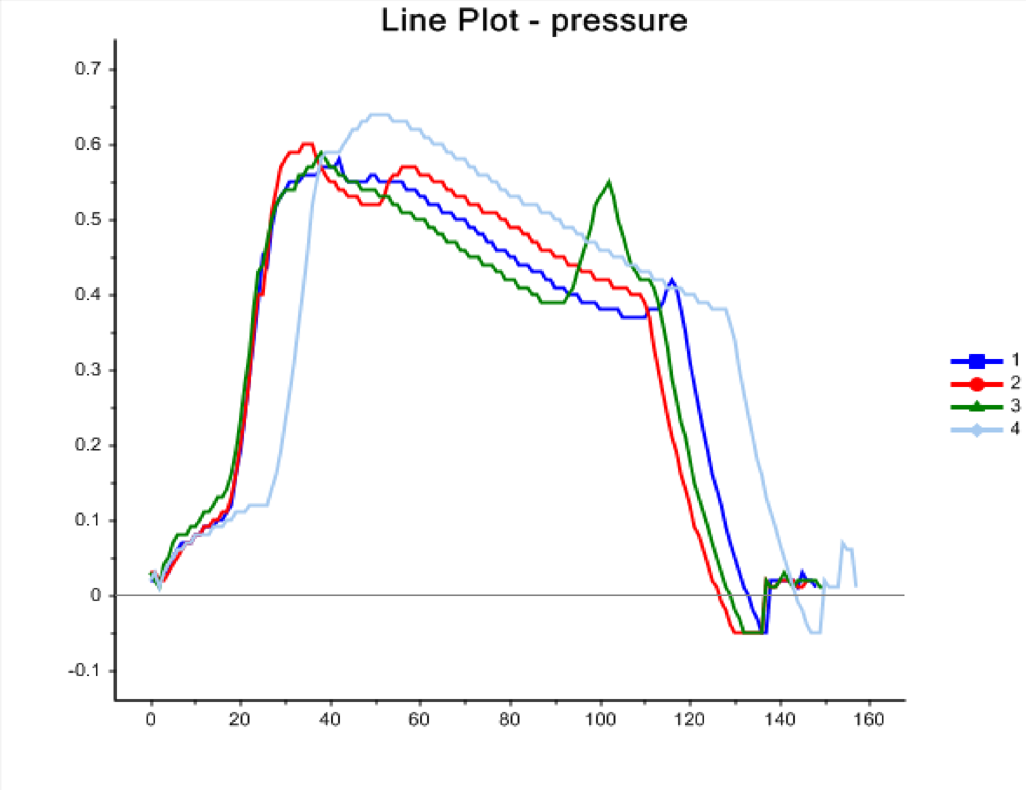
Pressure show in relative time

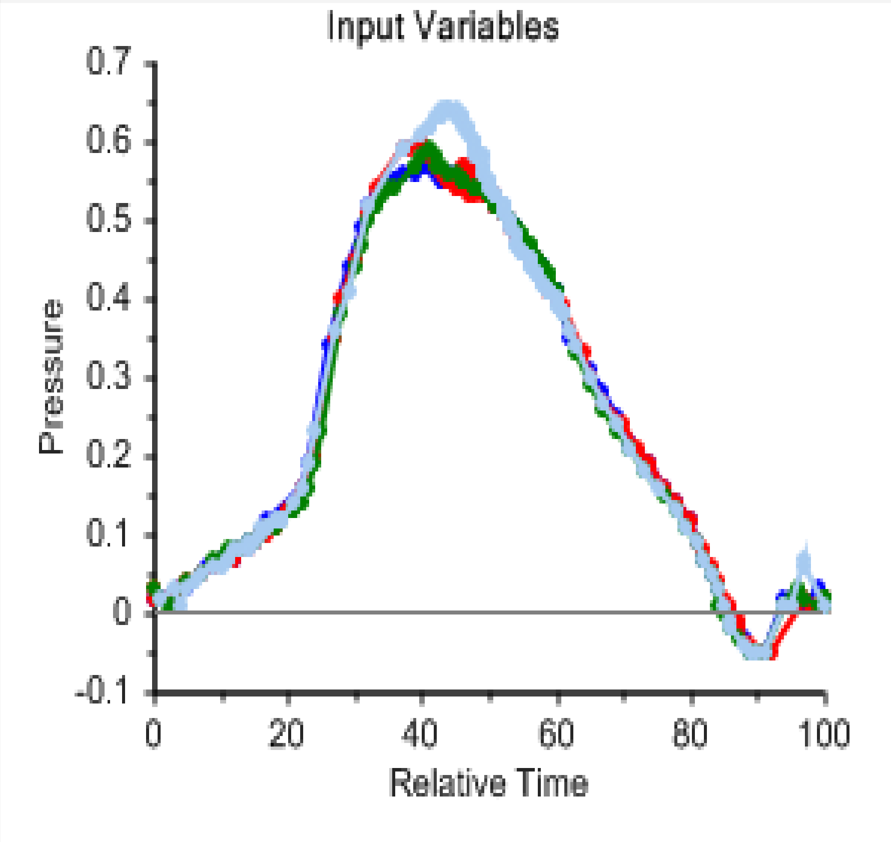


Comparison of the two

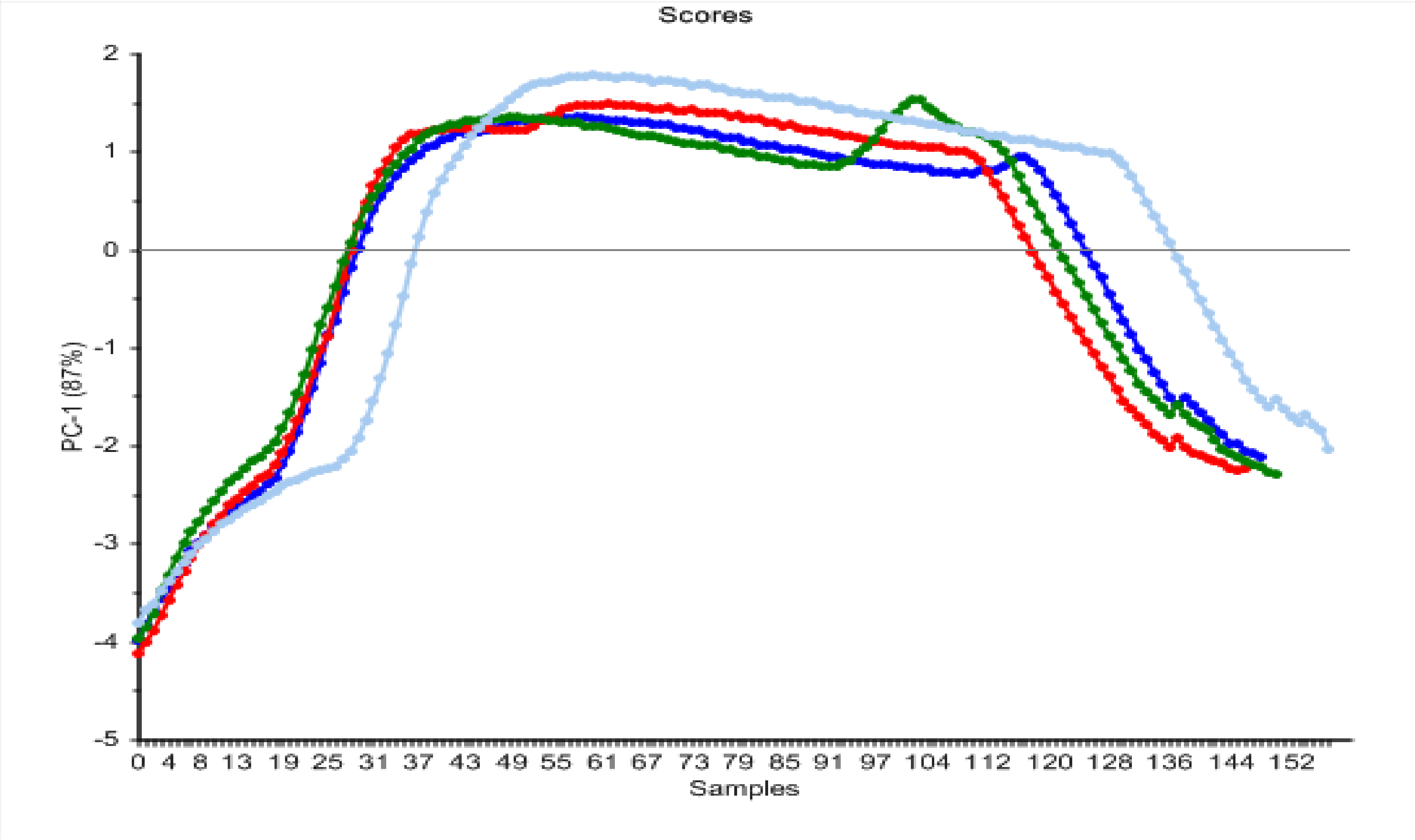
Looks like the batches are different

... but in reality: The same trajectory

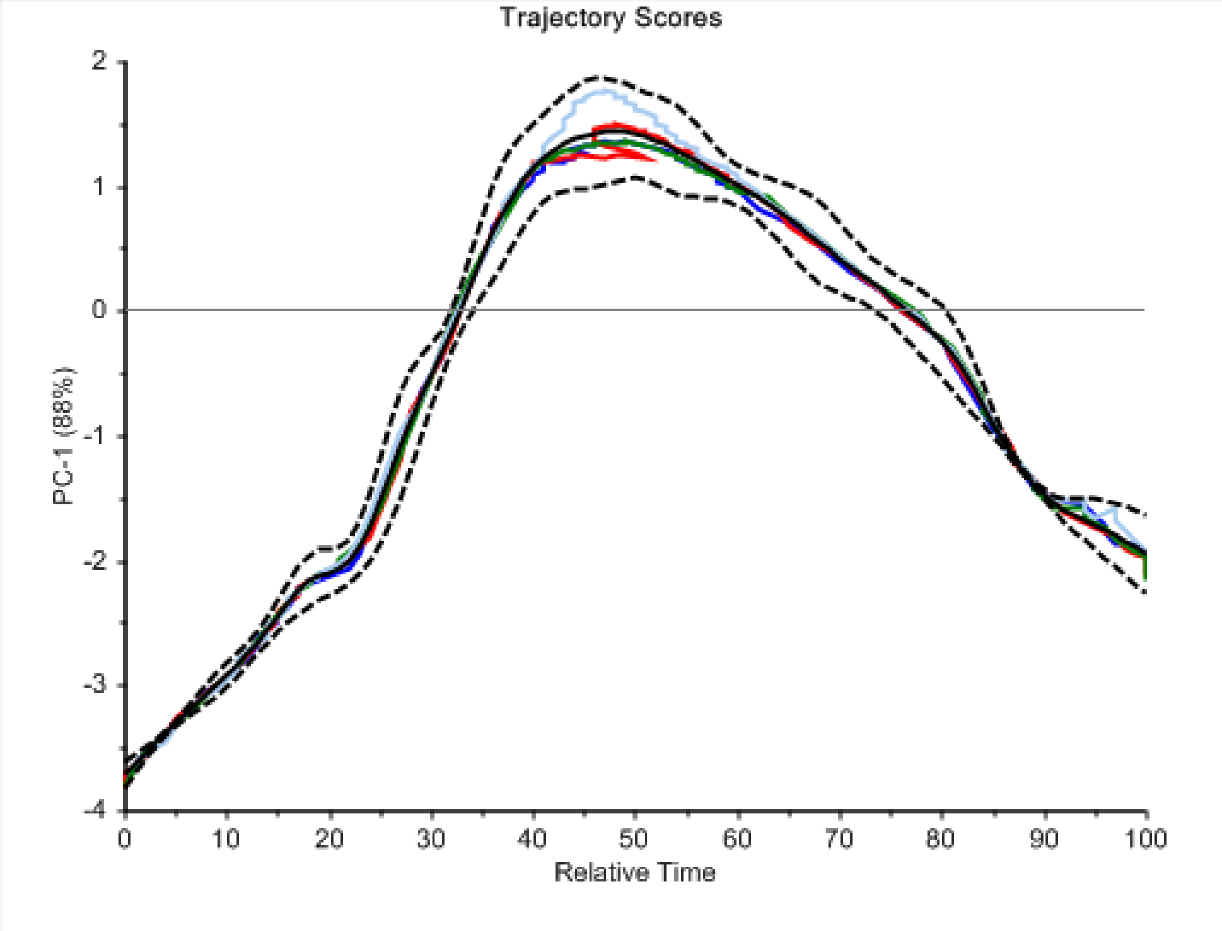




The scores showed as sample no.

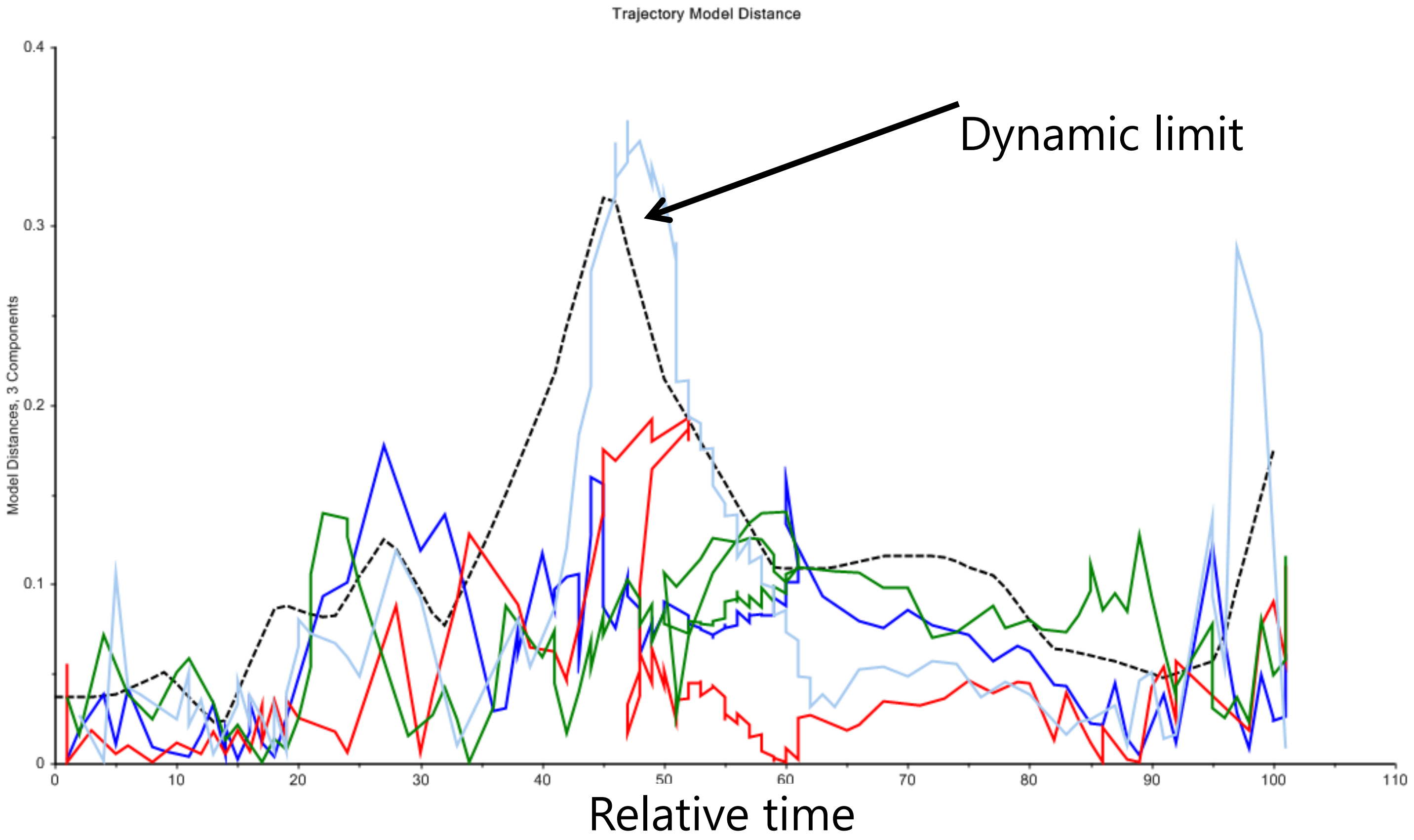


The scores showed in relative time



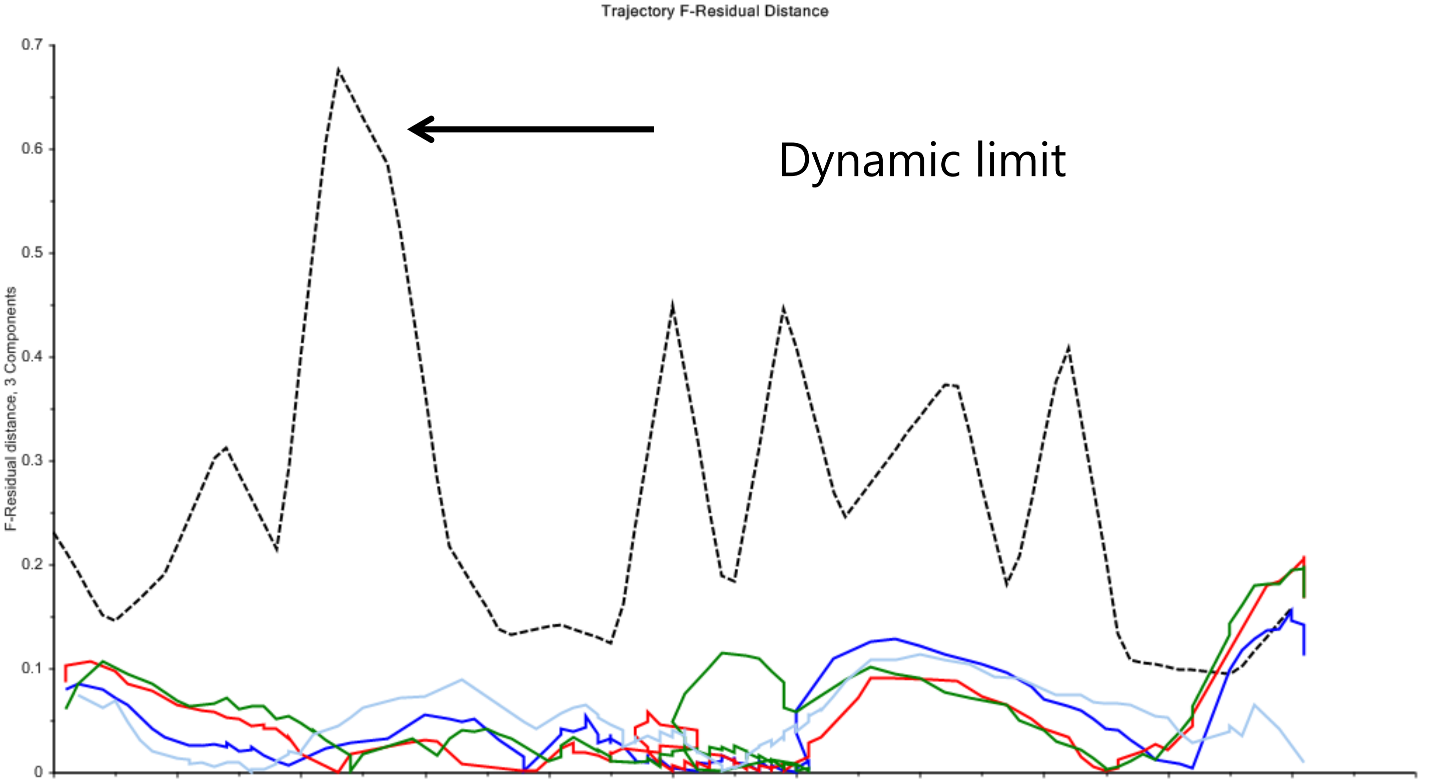
Distance to model in model space

Limits and historical batches



F-residual distance

Limits and historical batches



Dynamic limit

Relative time

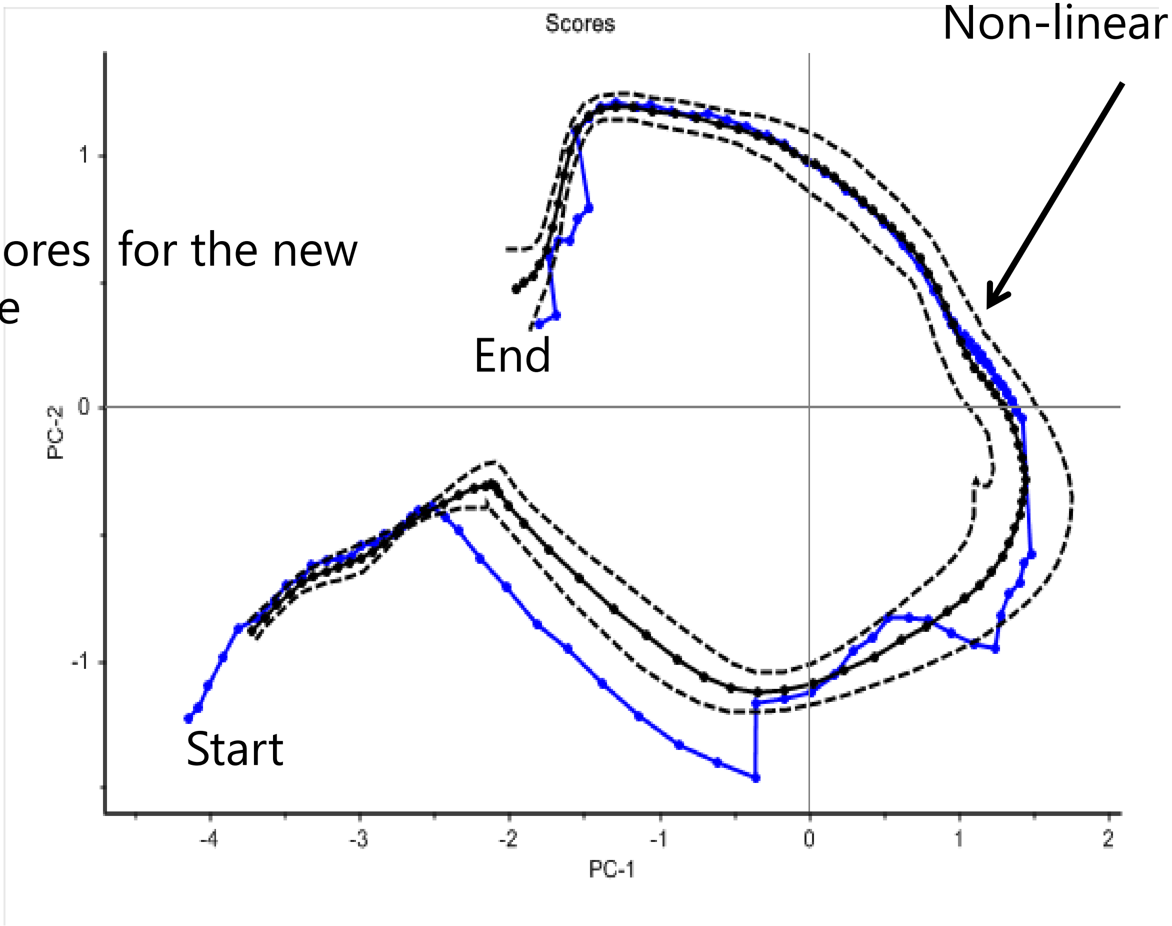
Chemical reaction

Average trajectory and 95% confidence interval

New batch projected

Non-linear behaviour

Projected scores for the new batch in blue



End

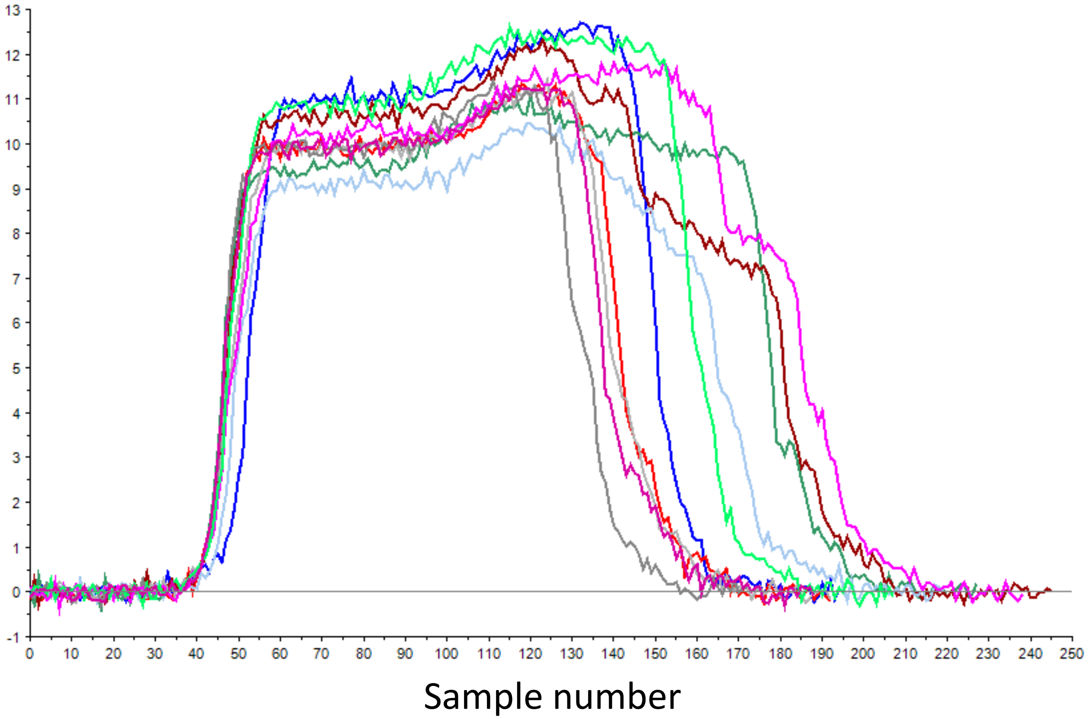
Start

Simulated fermentation data

- Fermentation data simulated with mechanistic models for production of penicillin, 10 for modelling, 3 for projection
- Process variables are collected on an hourly basis
  - Glucose concentration
  - Pyruvate concentration
  - Acetald concentration
  - Acetate concentration
  - Ethanol concentration
  - Biomass concentration (dry weight)
  - Active cell material
  - Acetaldehyde dehydrogenase
  - Specific oxygen uptake rate
  - Specific carbon dioxide evolution rate
- End-product quality: Biomass

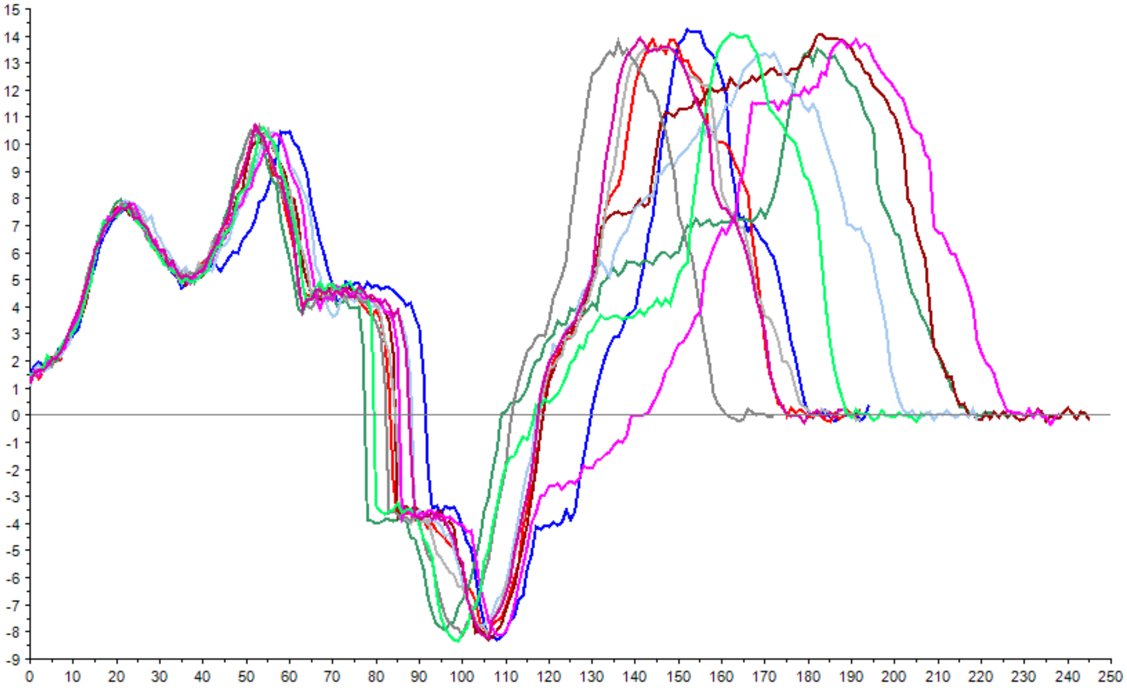
Line plot of Ethanol concentration

Grouped after batch ID



Line plot of Specific oxygen uptake

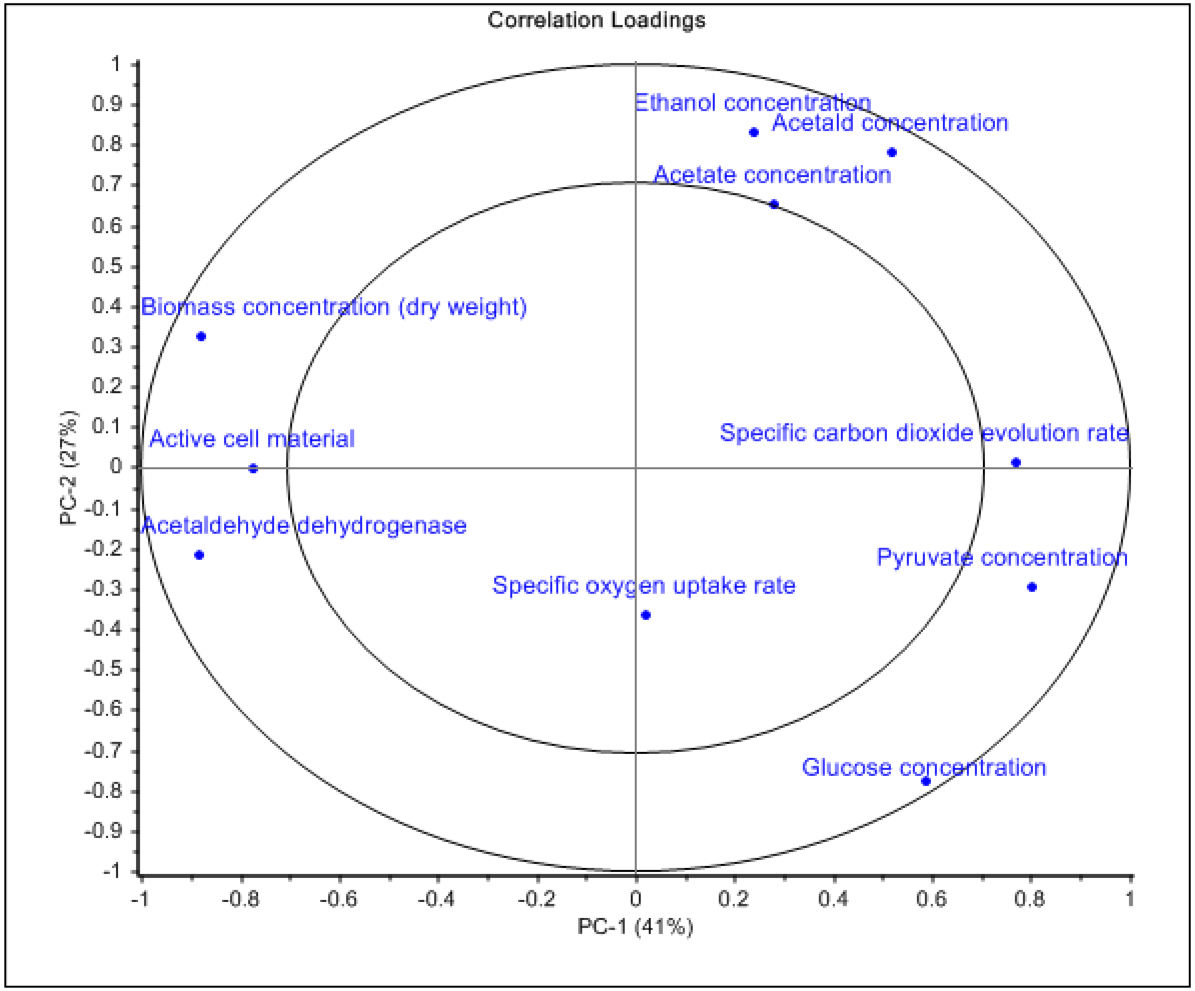
Grouped after batch ID



Sample number

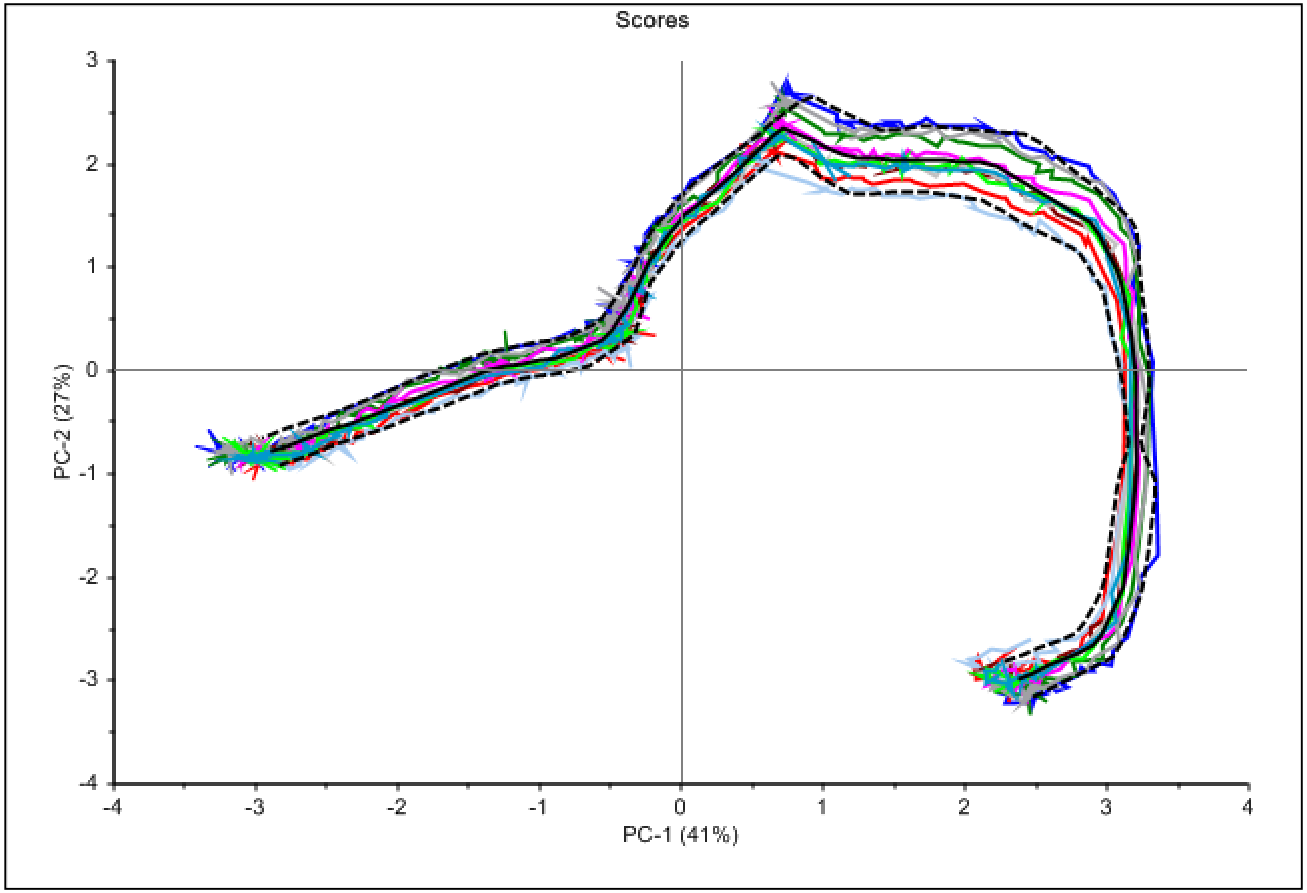
Results – map of variables

Correlation Loadings



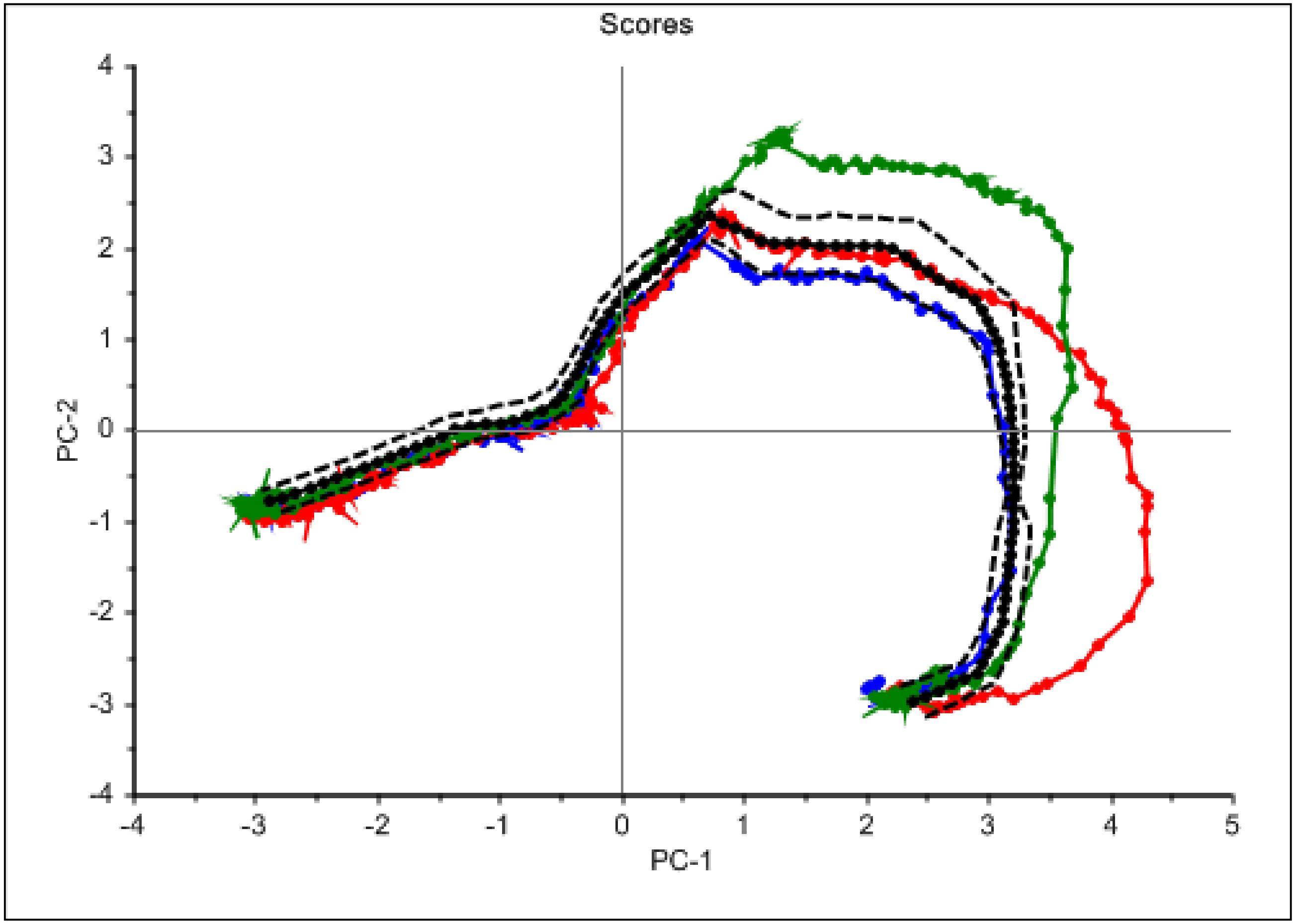
Results – map of samples

Score plot with batches, trajectory and limits



Results - projection

The normal batch in blue, the two with induced changes in red and green



Relative time – other applications

- Medicine: Monitoring patients as a function of the actual state of an illness (independent of time of consultation)
- Oil production: Simultaneous monitoring of the maturity of wells (production volume, water cut)
- Condition monitoring: Relative state of mechanical parts (wear & tear)
- … and more

Conclusions – Batch modelling

- This approach models the batch progression in relative time
- There is no need to:
  - Force the batches to a common length
  - Warp individual variables
- The method has a specific advantage in the monitoring phase as no attempt to synchronise is needed
- A one-dimensional representation of the batch's trajectory with dynamic confidence intervals is a compact way to visualise for operators

Conclusions

- Multivariate methods are already used in many industries, e.g. for prediction of raw materials and product quality
- Selectivity is not needed to make good predictions or for MSPC
- Redundancy among sensors makes it easier to detect sensor failure
- Humans should always be a part of model development and interpretation: non-artificial intelligence