Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #5

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/

Data challenges

- The main challenges on data are related to:
  1. **variability**:
     - systematic part of the signals should be distinguished from the noise
     - systematic variability can be introduced changing some factors of the system/process, for example using Design of Experiments (DoE)
     - the presence of noise should be considered to avoid drawing misleading conclusion
  2. **complexity**:
     - **a system is incomprehensible if the number of measured variables is >3/4**
       - simple statistics and graphical representations are not effective with multivariate datasets
  3. nature: data types can be categorized in several manners:
     - factors and responses
     - quantitative and qualitative
       - quantitative may assume any reasonable real value in a continuous scale
       - qualitative are categorical variables that assumed predetermined levels
     - controlled and uncontrolled
       - controlled variables can be manipulated, set to a determined value and kept there
       - uncontrolled variables are impossible to regulate, but may impact on the system/process

Today's lesson

- Let's go multivariate! :)
  - correlation/covariance
  - dimensionality reduction
  - latent variable modelling

    =

    projection modelling

Multivariate statistics

Multivariate vs. univariate data

- **Univariate data**:
  - can be easily studied through univariate statistics
    - analysis of the distribution
      - central tendency
      - variability
      - shape
    - probability theory: detection of less frequent and anomalous values
  - a lot of methodologies for data analysis and interpretation are available

- What if the **number of variables is higher than 3/4**?
  - is it practical to inspect variables one at a time?
  - is it appropriate to inspect variables one at a time?
  - is it sufficient to inspect variables one at a time?

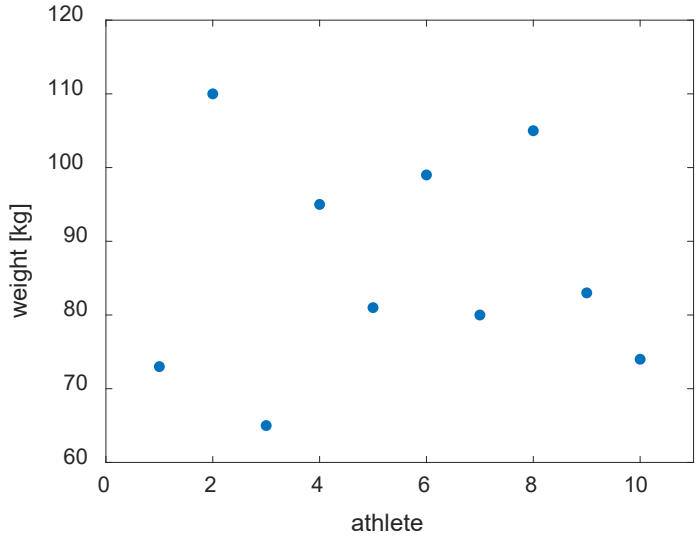
Example 1: physical condition of the athletes

… my first job: volleyball trainer!

Athletes weight

- Examining the body size of the athletes
- Weight data seems to be “standard”:
  - mean 86.5 kg
  - standard deviation 14.9 kg

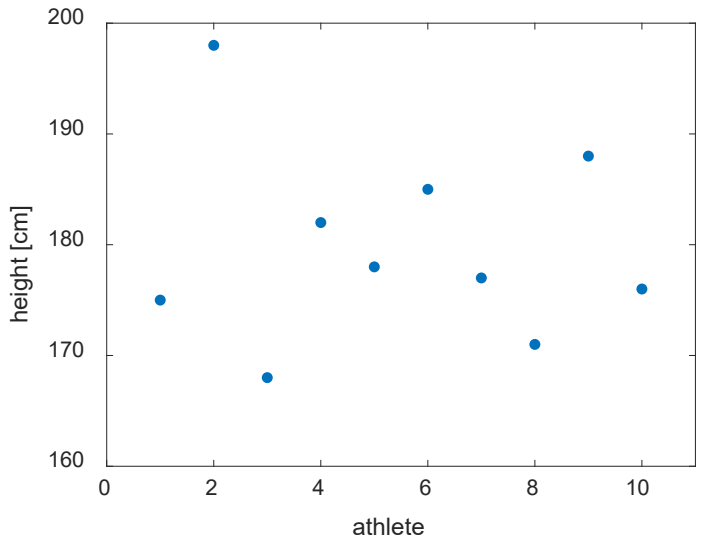
athlete	weight (kg)
1	73
2	110
3	65
4	95
5	81
6	99
7	80
8	105
9	83
10	74



Athletes height

- Also height data seems to be “standard”:
  - mean 179.8 cm
  - standard deviation 8.8 cm

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

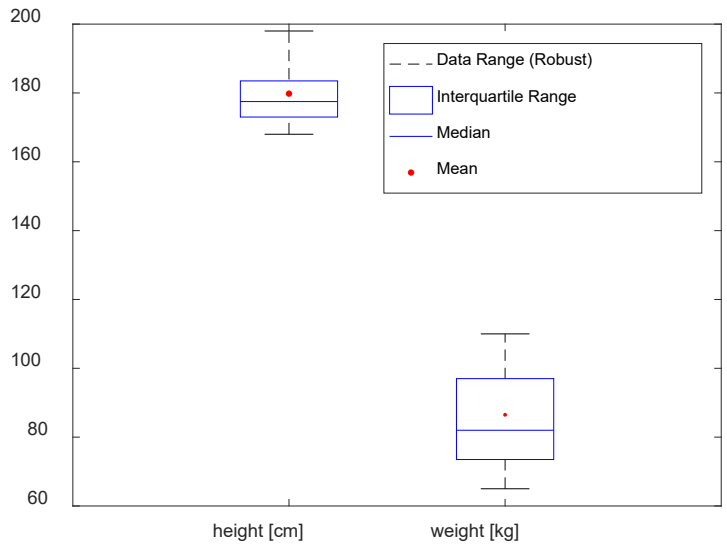


Inferred univariate statistics for athletes

- Sample statistics can be easily calculated

	height [cm]	weight [kg]
mean	179.8	86.5
st.d.	8.79	14.94
skewness	0.73	0.25
kurtosis	2.92	1.80

- … and also easily displayed

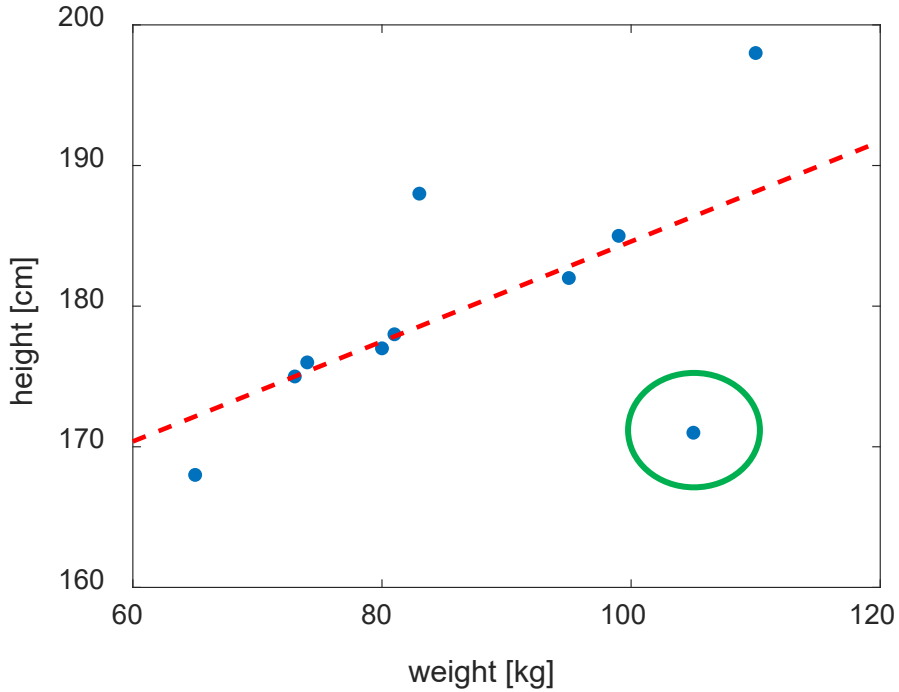


Is it practical?
Is it appropriate?
Is it sufficient?

Multivariate data observation

- Is there any additional information observing the data from the multivariate point of view (i.e., all together)?
  - could you find extra information in the joint view of the variables?
  - do you see anomalies?

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



Example 2: tablets quality

- Not only the number of the variables, but also the hidden content of information is valuable

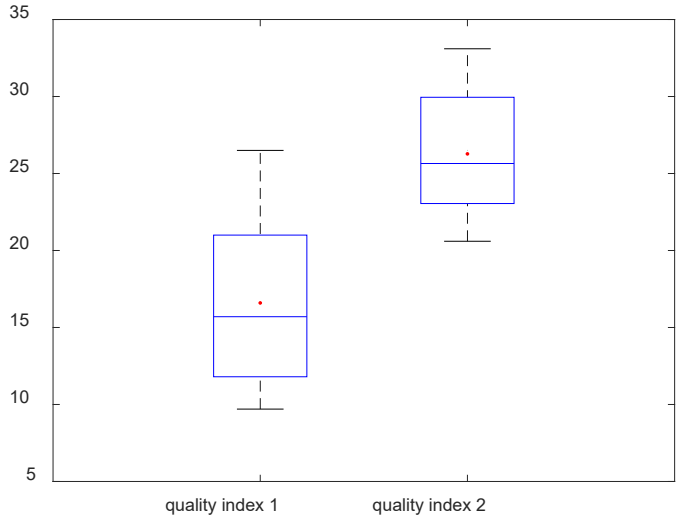
sample	quality index 1	quality index 2
1	21.2	32.5
2	16.2	21.0
3	13.1	21.7
4	11.6	21.3
5	20.8	29.9
6	10.4	20.6
7	19.5	26.8
8	9.8	25.2
9	15.2	31.2
10	12.0	26.0
11	17.6	28.5
12	24.0	30.0
13	17.8	33.1
14	15.0	24.0
15	11.0	24.2
16	24.8	25.3
17	12.8	23.3
18	26.5	30.6
19	22.9	27.5
20	9.7	22.8

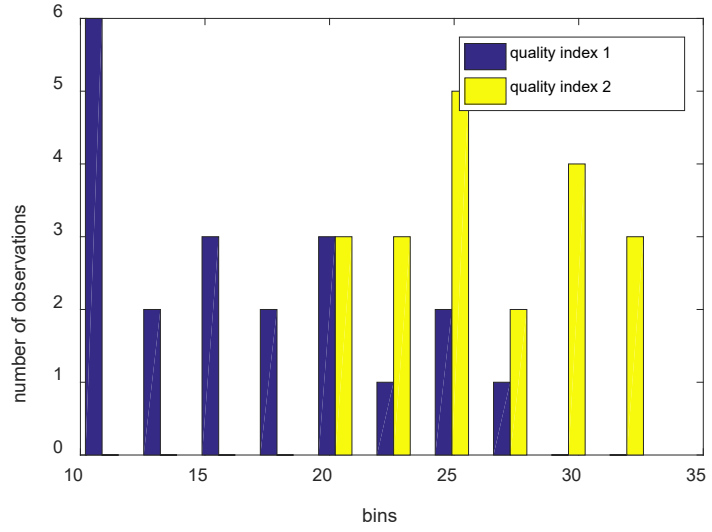
Inferred univariate statistics for tablets

- Sample statistics can be easily calculated

	quality index 1	quality index 2
mean	16.61	26.28
st.d.	5.38	3.98
skewness	0.35	0.20
kurtosis	1.84	1.80

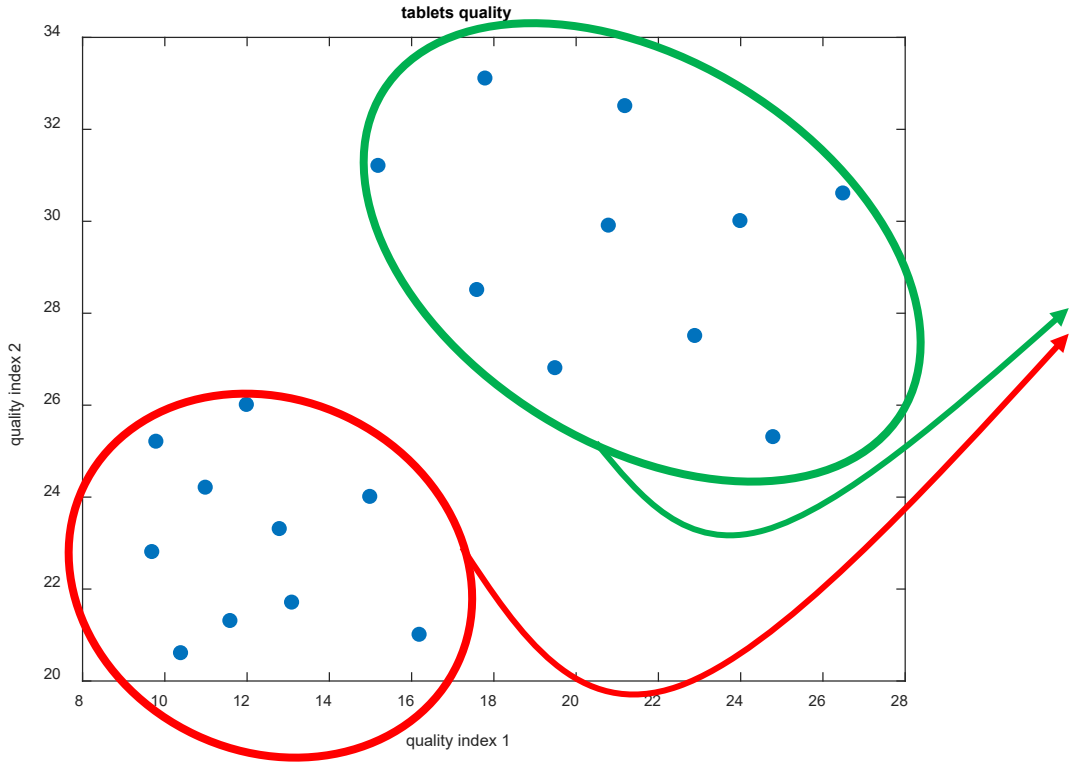
- … and also easily displayed





Drug quality in the pharmaceutical industry

- Multivariate data analysis shows different product classes
  - two different classes are identified



Different manufacturing periods:
- different product quality in mean and standard deviation
- the classes were identified despite the same process parameters were used to manufacture the tablets

Joint view on variables

- In multivariate datasets variables can be studied through the multivariate version of the parameters which describe:
  - location
  - dispersion

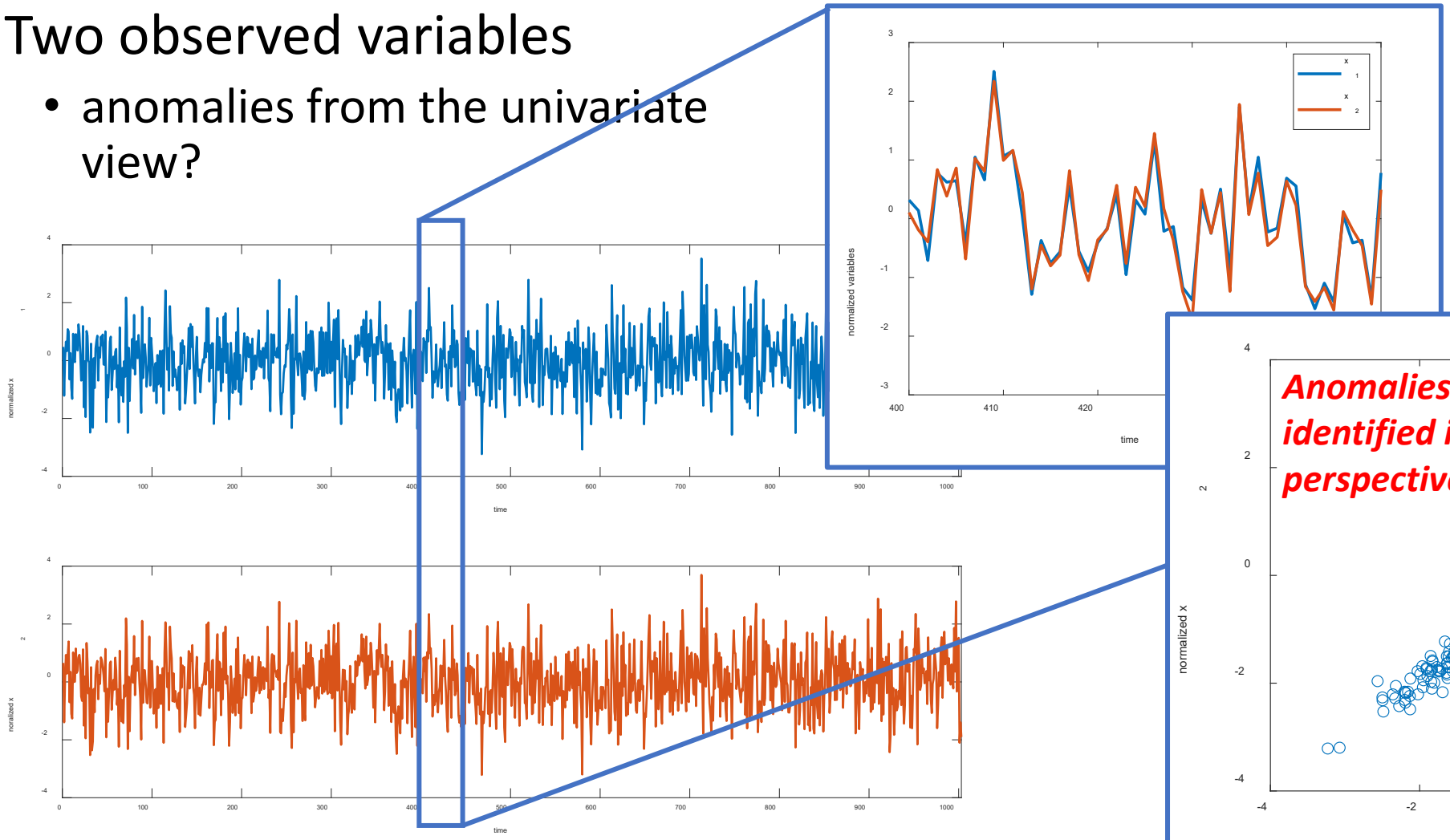
- **Location** is identified by the multivariate version of the mean of the variables $[\mathbf{x}_1 \quad \mathbf{x}_2]$:

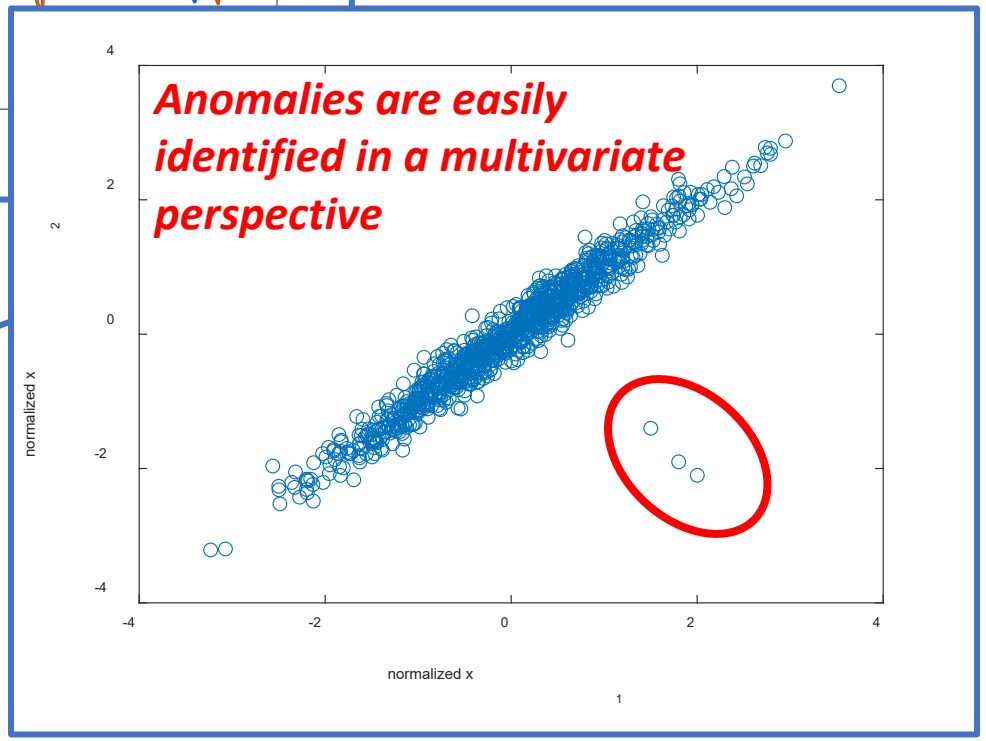
$$\boldsymbol{\mu} = [\mu_1 \quad \mu_2]$$

- Is the **variability** simply characterized by the variance of the single variable σ_1^2 and σ_2^2? Or is there something extra?

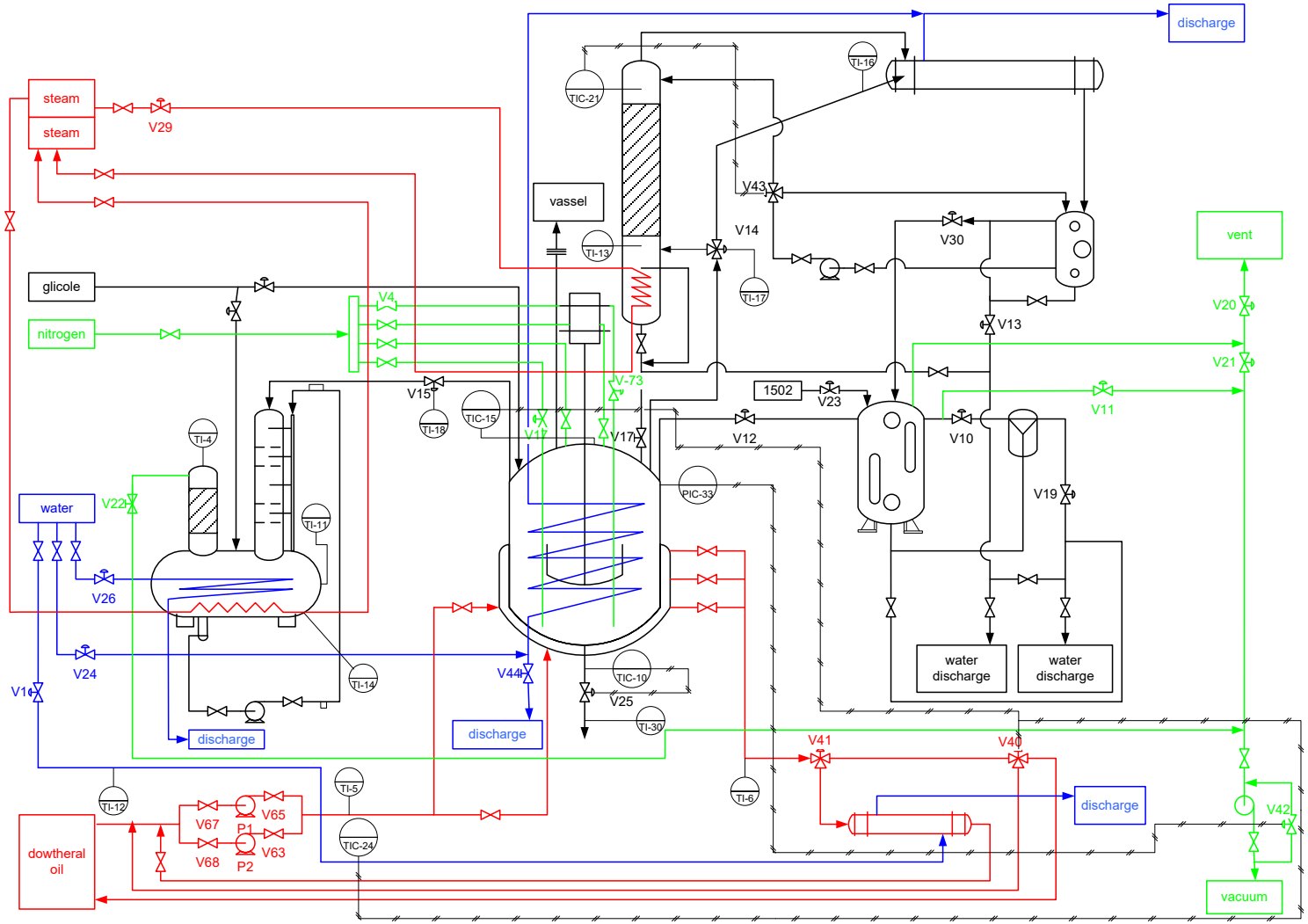
Example 3: observing anomalies in a process

- Two observed variables
  - anomalies from the univariate view?



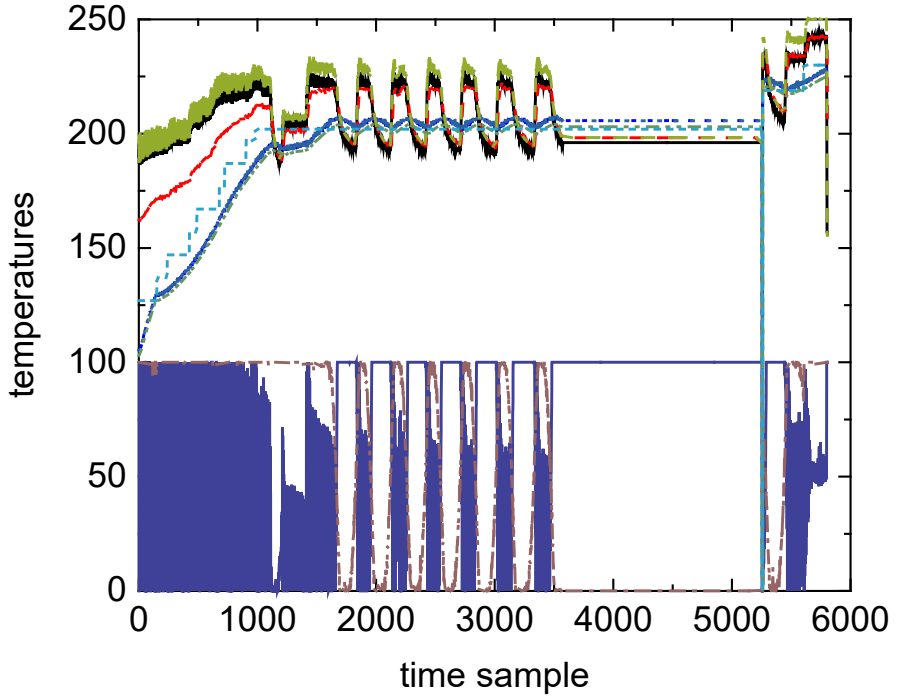


Example 4: production of resins and coatings



Time profiles of on-line measured variables

- Do you notice some evidence in the time profiles of these data?



- Multivariate data are often redundant:
  - a lot of **correlation** among different variables is present in the data

Correlation

- From the mathematical point of view, the **correlation** among two variables x_1 and x_2 is:

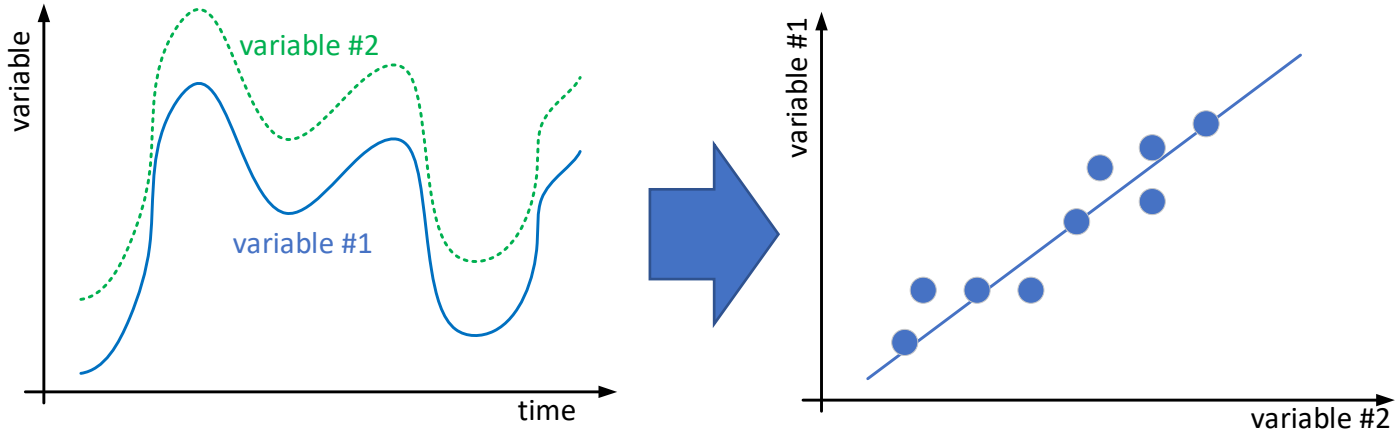
$$\rho_{x_1,x_2} = \frac{\sigma_{x_1,x_2}}{\sigma_{x_1}\sigma_{x_2}} \in [-1,1]$$

$$\begin{cases} \sigma_{x_1,x_2} = \frac{1}{N}\sum_{n=1}^{N}\left(x_{1,n} - \mu_{x_1}\right)\left(x_{2,n} - \mu_{x_2}\right) \text{ covariance} \\ \sigma_{x_1} = \frac{1}{\sqrt{N-1}}\sqrt{\sum_{n=1}^{N}\left(x_{1,n} - \mu_{x_1}\right)^2} \\ \sigma_{x_2} = \frac{1}{\sqrt{N-1}}\sqrt{\sum_{n=1}^{N}\left(x_{2,n} - \mu_{x_2}\right)^2} \end{cases}$$

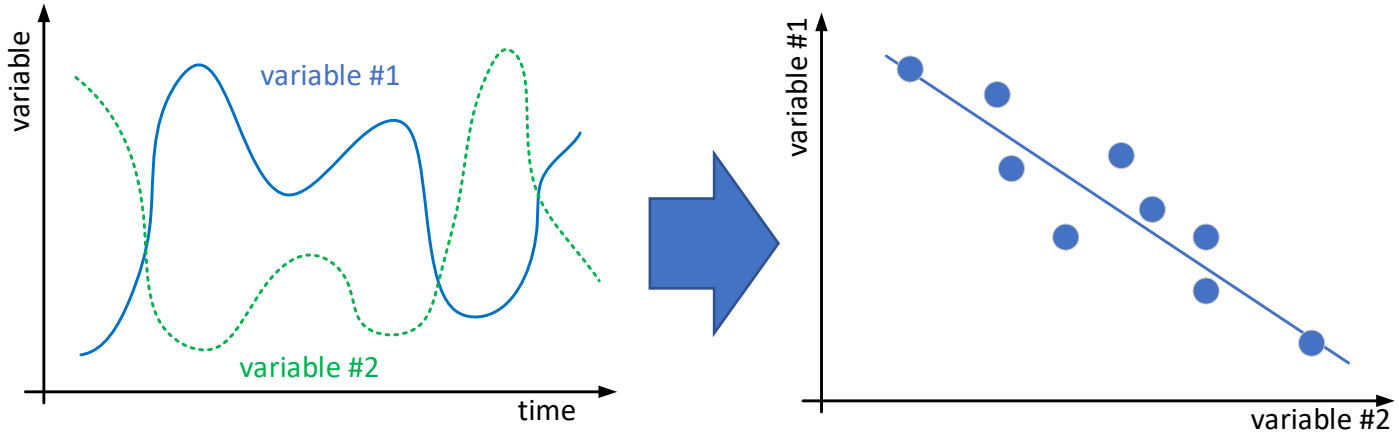
- Evaluating the **correlation structure** in a dataset means observing if data:
  - vary one in strict relation with the others
  - show common behavior (i.e.: trends, shape, etc…)
- In the example of athletes' height-weight:
  - *"standard"* athletes: the weight increases with height
    - the tallest persons are the ones with the highest weight
    - the shortest persons show the lowest weight

Correlation/covariance in practice (1/2)

- **Positive and large correlation**
  $\rho_{x_1,x_2} \sim 1$: two variables are positively correlated when they covary
  - when one goes up, also the other goes up
  - when one goes down, also the other goes down



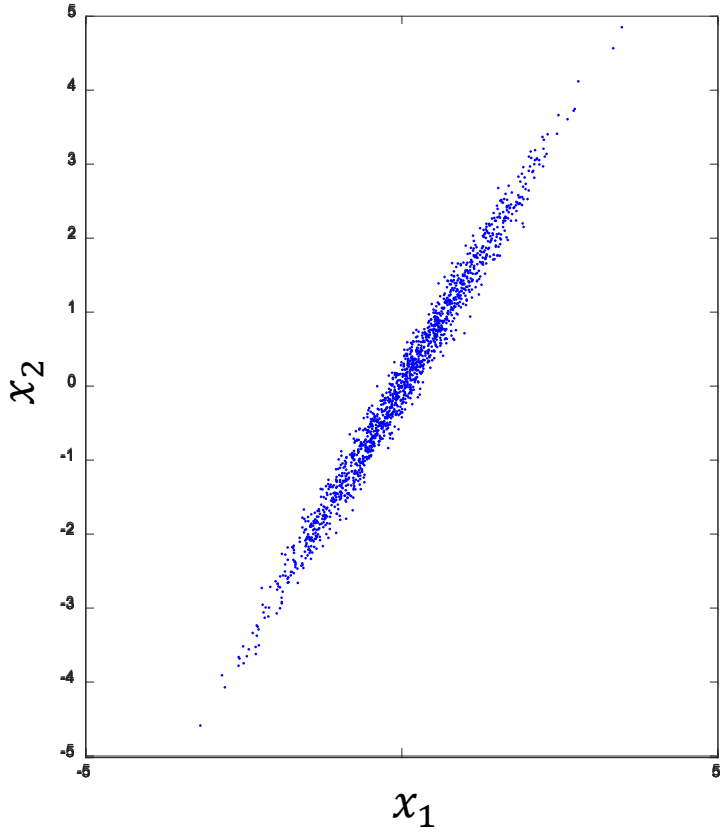
- **Negative and large correlation**
  $\rho_{x_1,x_2} \sim -1$: two variables are negatively correlated when they vary in opposite sides
  - when one goes up, the other goes down and vice-versa

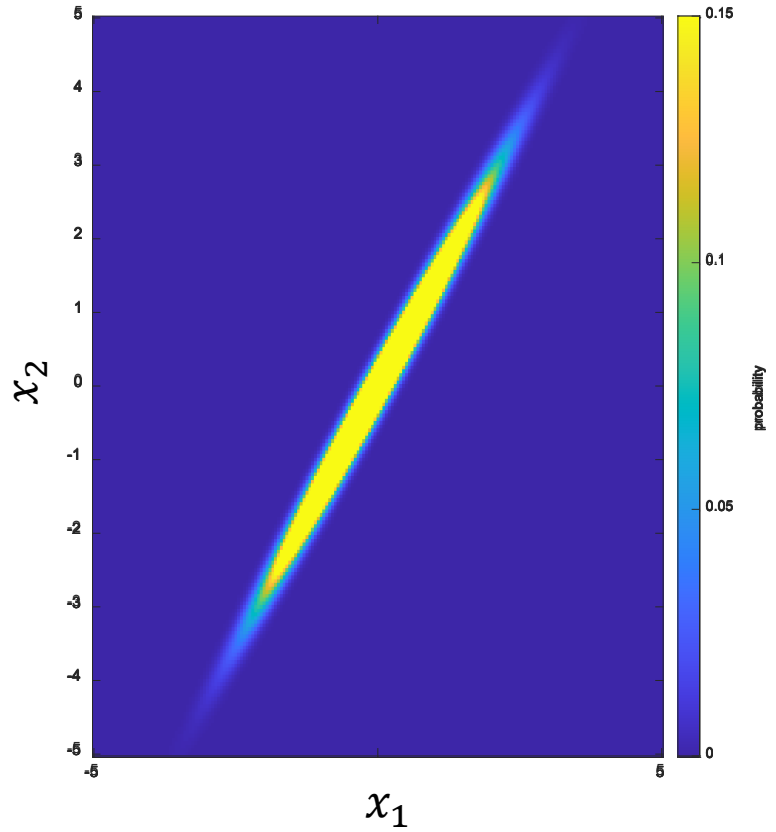


Correlation/covariance in practice (2/2)

$$\mu_{x_1} = 0;\ \mu_{x_2} = 0 \Rightarrow \boldsymbol{\mu} = [0 \quad 0];\quad \sigma_{x_1} = 1;\ \sigma_{x_2} = 2 \Rightarrow \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{x_1,x_2} \\ \sigma_{x_1,x_2} & 2 \end{bmatrix}$$

$$\sigma_{x_1,x_2} = 1.85$$





What misses in the univariate thinking?

- Joint view of all the variables **together**
- The ability of **visualizing** pattern
  - more than 3 **dimensions**
  - behavior of the data
- **Summarizing** the information of (a lot of) data and **interpreting** their information
- Dealing with data **correlation**
  - and also understanding **how they co-vary**

Multivariate version of the data

- The multivariate version of the data summary is composed of:
  - **location** is identified by the multivariate version of the mean:
    - for 2 variables only the mean is the vector:

$$\boldsymbol{\mu} = [\mu_{x_1} \quad \mu_{x_2}]$$

  - **dispersion** is composed of:
    - variance of the single variables σ_x^2 and σ_y^2
    - **covariance** σ_{x_1,x_2}

    … which are collected in the matrix $\boldsymbol{\Sigma}$ that for 2 variables only is:

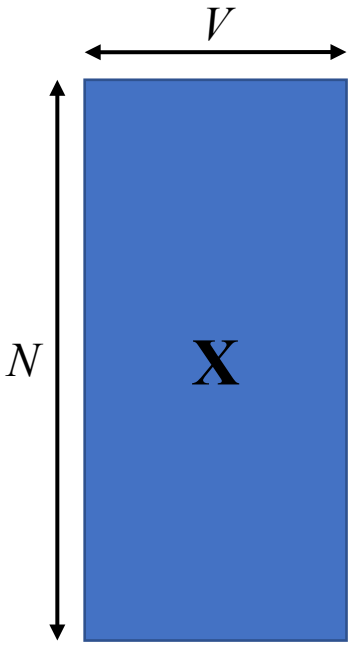
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1,x_2} \\ \sigma_{x_1,x_2} & \sigma_{x_2}^2 \end{bmatrix}$$

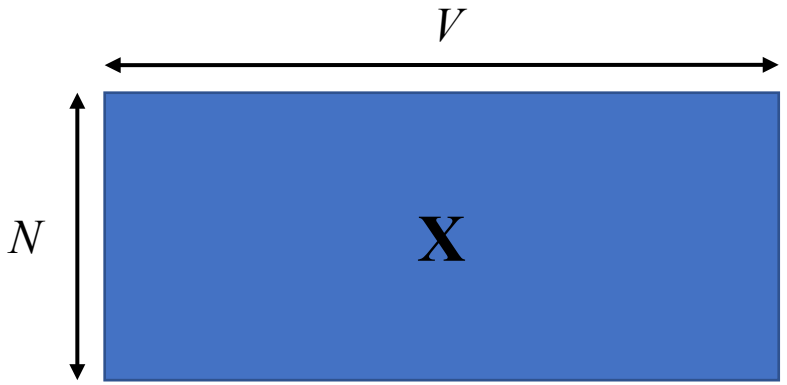
Multivariate data challenges

- The main challenges of multivariate datasets are related to:
  1. **dimensionality**:
     - **thousands of variables are recorded every few seconds** thanks to the instrumental revolution (instinctively we know that the more we want to understand, the higher the measurements number we need)
     - all data points are needed for a **proper inspection**
       - do not discard variables or samples if there is not a strong motivation!
     - not always correct to refer to some “reference” variables
  2. **multi-collinearity**:
     - variables are usually **correlated** one another
       - **collinear variables** are (approximately linear) function of other variables
     - information can be found in the correlation pattern rather than in the individual signal
     - although thousands of variables are available often only few underlying (latent) phenomena affect the system/process
     - the interpretation of the relation between variables is not straightforward
  3. **noise**:
     - disturbing factors introduce unwanted (known or unknown) variability
     - important effects may be partially obscured by noise
  4. **missing data**:
     - data tables may be partially incomplete and missing data are usually present in data historians (i.e., sensor failures, transducer problems, etc…)

Wide validity of the multivariate methods

- The abovementioned considerations are valid for both:
  - **tall arrays**: high number of observations N
    - typical of continuous processes
  - **wide arrays**: high number of measured variables V
    - usually available when adding observations (increasing N) is too time- and cost-intensive





Take-home message

«…having a lot of good data
without knowing how to interpret them
is like playing a grand-piano with only one finger!»

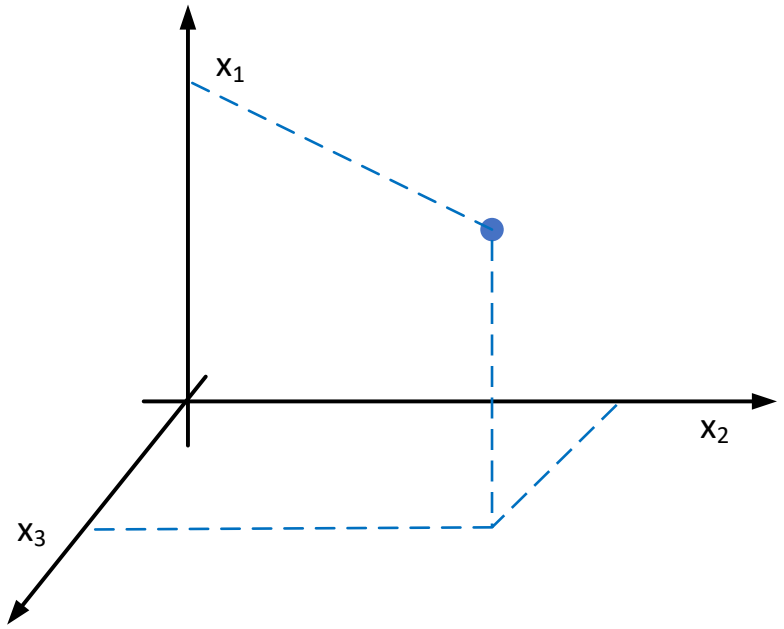
Start thinking multivariate!

Multivariate statistical techniques

- Multivariate statistical techniques are appropriate tools to face multivariate data challenges:
  - **dimensionality reduction**: they usually are projection methods that **compress the data dimensionality** from the original multidimensional space of the V variables to a much-reduced **space of latent variables** that represent the **physical phenomena that affect the system/process**
    - **informative diagnostics** and **effective graphical tools** can be provided
  - **correlative methods**: they are correlative methods to deal with correlation and to compress original variables in latent variables
  - **noise filtering**: the non-systematic part of the signals can be easily discarded and removed from the main effects influencing the system/process
  - **missing data handling**: limited amounts of missing data (10-20%) are tolerated without affecting the robustness of the models

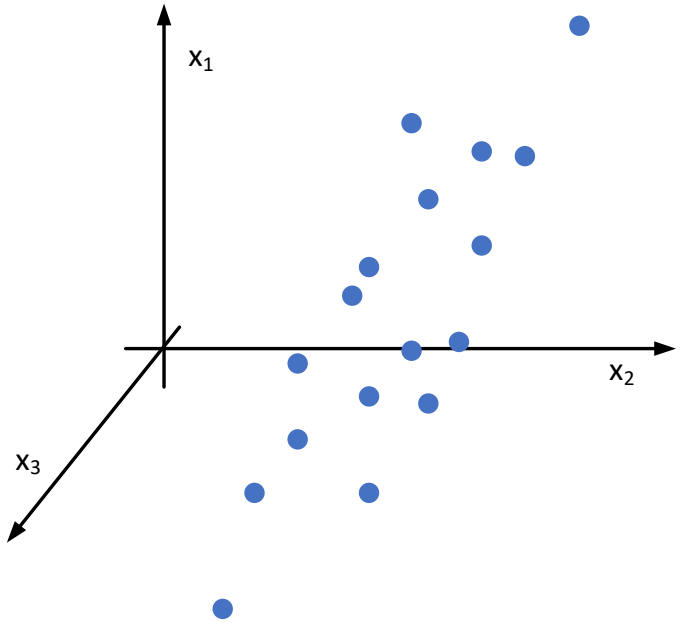
Projection-based models (1/4)

- Consider $\mathbf{x}_n$, one observation in the V dimensional space of the original variables (the n-th row in $\mathbf{X}$) with $V = 3$ for the sake of simplicity and visualization
  - it is the array $\mathbf{x}_n = [x_{n,1} \quad x_{n,2} \quad x_{n,3}]$
  - the case can be easily extended to the case of $V \gg 3$



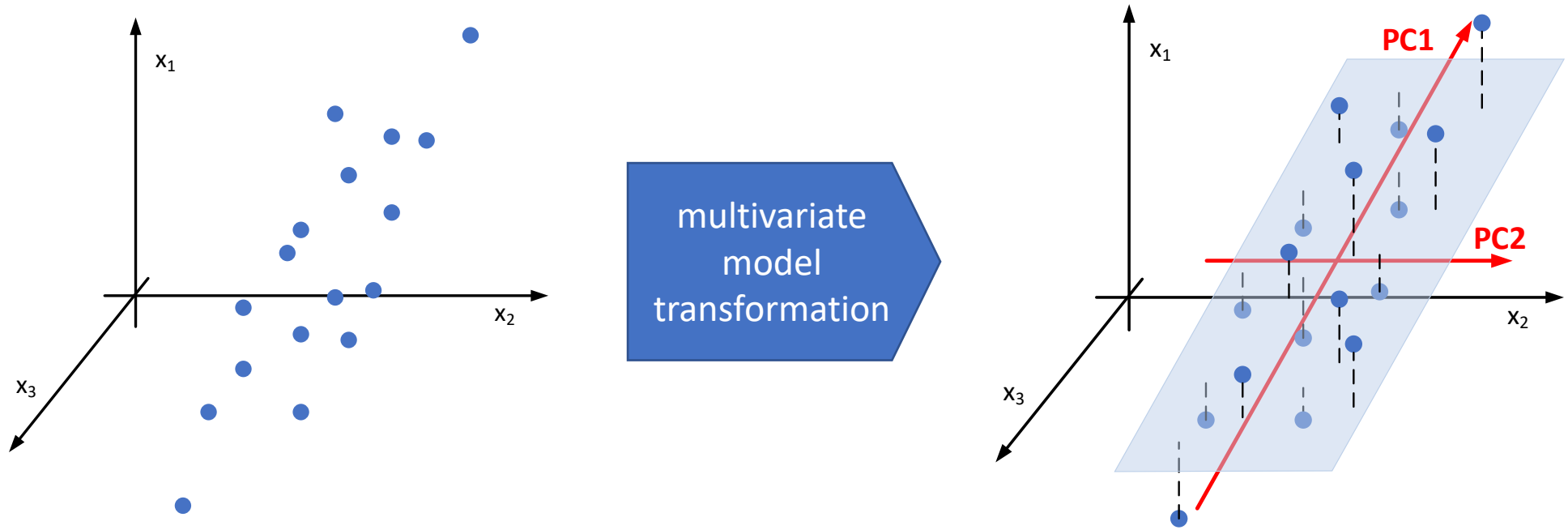
Projection-based models (2/4)

- The observations in matrix $\mathbf{X}$ can be displayed in the V-dimensional space as a swarm of points
- Analyzing the data in a multivariate perspective corresponds to formulating a *quantitative description of the shape* of this swarm:
  - the model should **approximate the data** in a sufficiently low-dimensional sub-sace to convert the data in an **informative and easy-to-understand (and to visualize) manner**



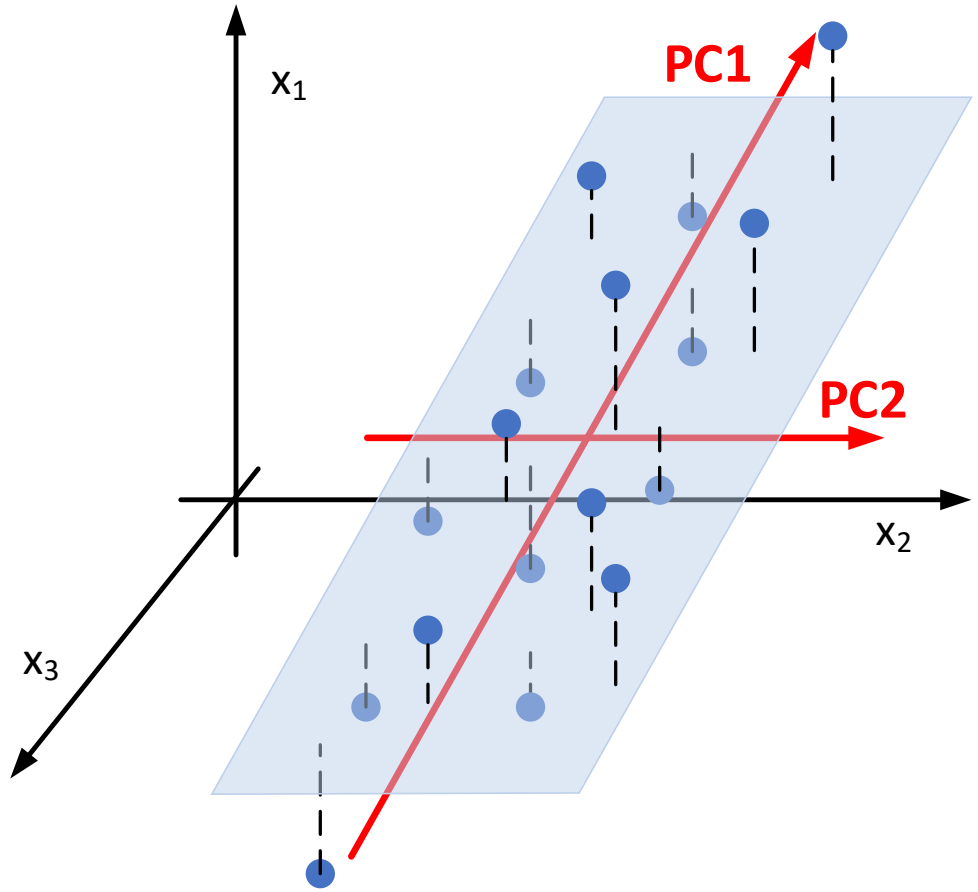
Projection-based models (3/4)

- When data are collinear, few latent directions can explain almost the entire information content stored into the data $\mathbf{X}$
- The shape of the data can be represented in this low dimensional space of latent directions where the original data are projected



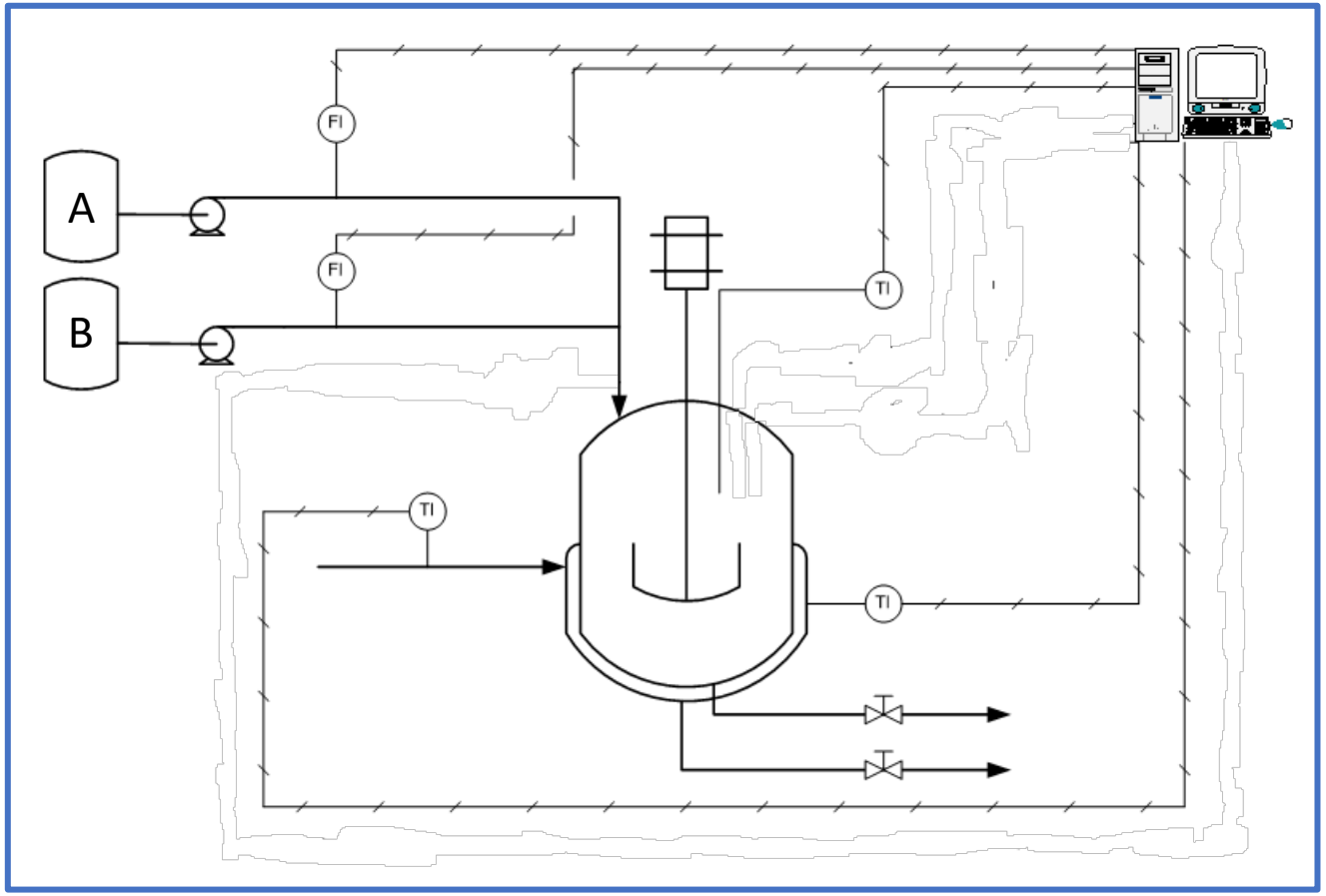
Projection-based models (4/4)

- The projection requires:
  - passing from a high-dimensional hyperspace of V dimensions to a much-lower hyperplane of $A << V$ latent directions
  - the identification of the **direction of maximum variability of the data**
    - this implies also to understand the correlation and to compress the dimensionality
    - the latent directions identify what variables are important
  - the **most influential observations** are the ones in the periphery
    - the less influential ones are located close to the origin
    - the most similar observations are closer in the space



Example: reactor for the production of rubber (1/3)

- Exothermic reaction between monomers A and B to manufacture rubbers
  - reaction depends on 2 independent phenomena:
    - material transfer & reaction
    - heat transfer
  - 5 (correlated) measurements are available:
    - reactor temperature, jacket temperature and cooling water temperature are related to heat exchange
    - raw material flowrates are related to material transfer



- What is the best way of summarizing these data?

Example: reactor for the production of rubber (2/3)

- Data are available in form of tables (i.e., matrices):

observation #	reactor T [°C]	jacket T [°C]	cooling water T [°C]	A flowrate [kg/h]	B flowrate [kg/h]
1	103.4	43.2	15.7	20.8	10.4
2	107.2	48.9	16.1	22.2	11.1
…	…	…	…	…	…
N	102.0	44.5	14.9	21.6	10.8

- A bidimensional space made of the 2 independent phenomena is sufficient to represent all the data
  - removing “redundant data” is not a good solution: data have a lot of information, why discarding it?

Example: reactor for the production of rubber (3/3)

- We can **summarize** the information of the original correlated variables x_i (i.e., reduce the dimensionality of the system from a 5-dimensional space of temperatures and flowrates) in a system of 2 dimensions whose coordinates are made of the **latent phenomena**, namely, new variables t_1 and t_2 obtained by weighted averages of the original ones x_i which highlight the independent phenomena, where the weights are $p_{i,j}$:
  - $t_1 = x_1 p_{1,1} + x_2 p_{1,2} + x_3 p_{1,3} + x_4 p_{1,4} + x_5 p_{1,5}$
  - $t_2 = x_1 p_{2,1} + x_2 p_{2,2} + x_3 p_{2,3} + x_4 p_{2,4} + x_5 p_{2,5}$

where:
- t_i are called **scores**
- $p_{i,j}$ are weights, called **loadings**
- t_1 represents the **heat transfer** ($p_{1,4} \sim p_{1,5} \sim 0$)
- t_2 represents the **feeding flowrates** ($p_{2,1} \sim p_{2,2} \sim p_{2,3} \sim 0$)

Theoretical foundations of the projection models

- Theoretical foundations:
  - **latent variables**:
    - the original variables are represented in terms of latent variables
    - latent variables are linear combinations of the V original variables
  - **Taylor expansion**:
    - the data $\mathbf{X}$ are supposed to be generated as functions $f(\mathbf{T}, \mathbf{P})$, where:
      - $\mathbf{T}$ describes the changes in the observations
      - $\mathbf{P}$ describes the changes in the variables
      - the smaller the interval of $\mathbf{T}$ that is considered, the fewer the terms in the Taylor expansion we need in the model
- The original data can be interpreted as:
  - **observations on linear combination of measurements**
  - **measurements on a set of relatively similar and linearly dependent observations**
  - a mixture of the previous
    - multi-linear behavior of the data

Today's homework: flipped learning procedure

- I propose to you a **flipped lecture**
  - the learning procedure is inverted
- Please, complete the following procedure before next lesson:
  1. **attend the video lecture #6 already** available in Moodle
  2. **read the following papers** (they are available among the "Suggested readings" of Moodle)
     - Geladi, P., Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta,* **185**, 1–17
     - Wise, B.M., Gallagher, N.B. (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control,* **6**, 329–348
  3. **prepare questions** and anything you need to discuss with the teacher and your mates in the following lecture
- Next lesson will be held in the following manner:
  - 45 min of Q&A:
    - questions (of the students) and answers (of the teacher)
  - 45 min for the dealing with an example

… per sempre a fianco a me!