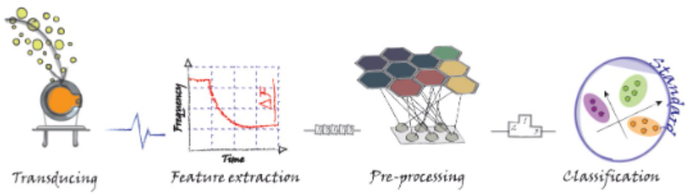


Lesson 2

Why could we start from the multivariate data analysis?

From univariate to multivariate data analysis

1

Univariate data

- Several analytical procedures produces an univariate data, where the experimental data depends by a single variable:
  - Measure of a single unknown variable.
  - Accurate control of the interferences.
  - Maintain constant the RH environmental conditions
  - ……
- An univariate data is defined as scalar value and a physical measure.
  - Example:
    - The measure of electronic resistance is 100KΩ
    - The weight of an apple is 80g
    - The concentration in the water of K^+ is 1.02 mg/l
- A sequence of measures results in a set of univariate data

2

2

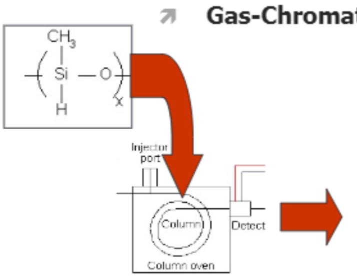
Univariate and multivariate data

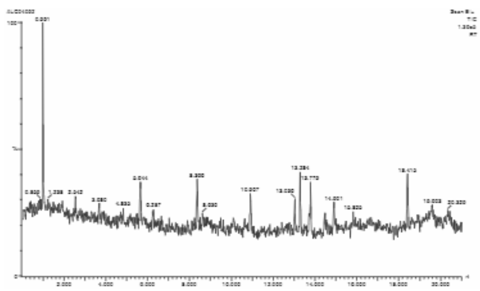
- Nevertheless a lot of phenomena are intrinsically complex and only considering several indicators it is possible to have a reliable model.
- Some instruments, as for example Gas Chromatography, spectrum analyser, produce as measurement result a multivariate data only for a single sample.

3

Instruments with multidimensional output data(1)

- **Gas-Chromatography**





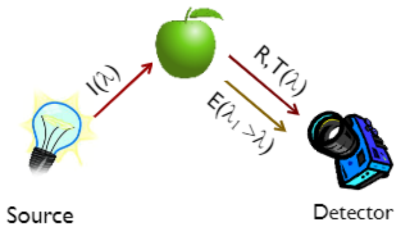
For each sample the instrument will give as output a spectrum (signal intensity vs. elution time)

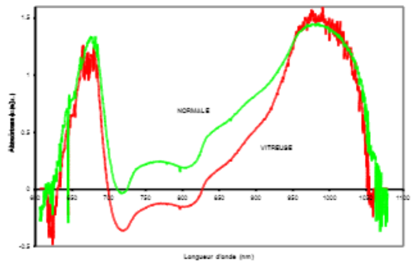
4

4

Instruments with multidimensional output data(2)

- **Spectroscopy**
  - Example: Spectroscopy Vis/NIR of a fruit.





5

5

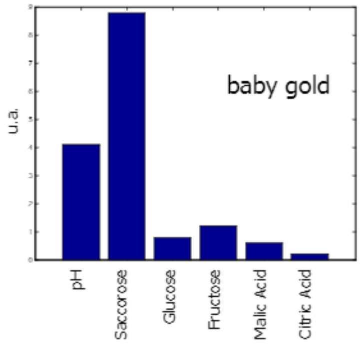
Instruments with multidimensional output data (3)

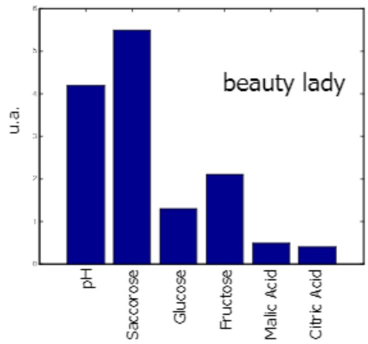
- **Description of complex phenomena**
  - Food quality (ex. fruit)
    - Sugars, Acid, pH, ethylene …
  - Mineral Waters
    - pH, CO_2, Cl, Na, K, Mg,…
  - Wheater conditions
    - T, RH, wind velocity,,…
  - Status of car
    - Velocity, accelerations, oil brakes, oil engines,…
- Each descriptor results, *in theory*, from an independent measurement.
- The set of descriptors describing a sample is called pattern.
- The operation that assigns a pattern to a specific class is called **Pattern Recognition.**

6

Instruments with multi-dimensional output data (4): Pattern

- **Sensor Array**
  - Example: biosensor set for the measure of pH, saccarore, glucose, fructose, malic acid, citric acid in peach and nectarines





7

Multivariate data

- ***We have also a multivariate data when a measure on a sample results a ordered sequence of scalar quantities***
  - ***The order is related to the physical meaning of the measure.***
- Source of multivariate data:
  - *Instruments or measurement techniques that intrinsically give multivariate data*
  - *The investigation of complex phenomena or samples requires the collection of several descriptors/measures in a multivariate data.*
- From the mathematically point of view a ordered sequence of number *N* is a vector.
- To each multivariate measure corresponds to a vector in a multidimensional space.

8

Data: which kind ?

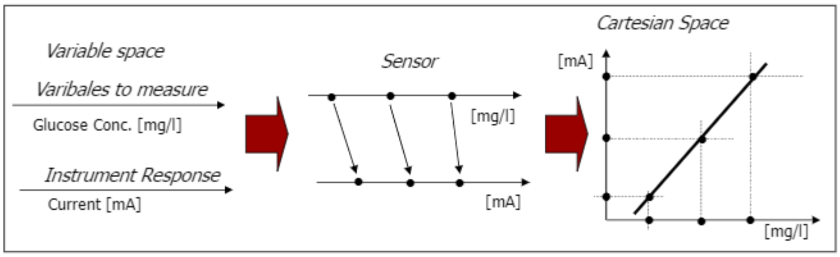
- **Quantitative (hard)**
  - It is a number and it has a physical dimension
    - *The temperature of the water is 400.0 K*
  - The quantitative data are the base of the galileian science and the *"hard sciences"* : disciplines based on mathematical modelling
- **Qualitative (soft)**
  - Labels, categories, ...
    - Ex: the water is cold, the shoes are large,...
  - Data that are difficult to standardize (sensorial analysis)
    - *Fuzzy logics*
- **Discrete data:**
  - Predefined value, discrete range
- **Continuous Data**
  - Limited range, continuous values

9

9

Fundamental criterion for data analysis

- The data can be represented in Euclidean vector space.
- Each observable is related to a dimension and correspond to a specific base vector.
- Reference system of the vector space is defined as a set (base) of orthonormal vector equals to the number of observables considered in the analysis
  - This is a straightforward for the univariate data:



10

Fundamental criterion for data analysis

For the multivariate data, it will observe that:
The core of data analysis consists in the definition of a new representation (basis) obtained as linear or non linear combination of the observables that maximizes the the information content of data.

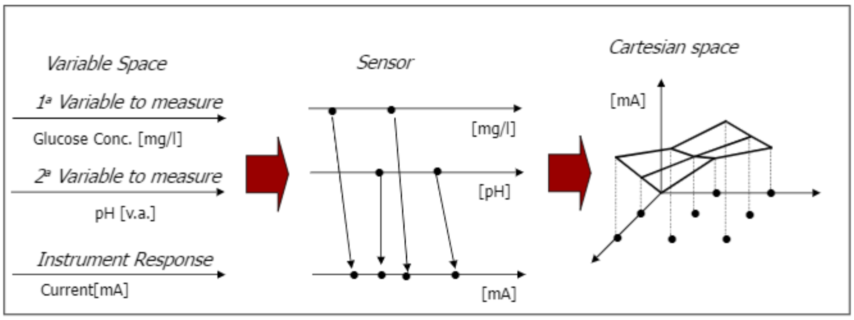
datum → *Analysis* → information → *Model* → Knowledge

11

Multivariate Data analysis

- The multivariate data analysis lets to connect between them the two vectors

$$y = f(x) \rightarrow \vec{y} = f(\vec{x})$$



12

12

General problem of multivariate data analysis

- Basically the main goal is to obtain from an instrumentational measurements information about the sample under investigation.
- In the univariate case (1-dimensional) the instrument gives an *"output"* for a single *"input"*

$$y = k \cdot x$$

y: out of instrument
x: input coming from the unknown sample
k: peculiar characteristic of the instrument

- **The main goal is, how I can obtain information of x starting from the output y and knowing k?**
  - With the Calibration!!
  - Calibrate the instrument means exposure to well-known strains (x) and measuring the output y, we can derive the value of k, and use it for a future and unknown sample.

$$x = \frac{y}{k}$$

13

13

Calibration

- Each single sensor is described by a characteristic function che put in relationship the output (V signal) with the input (measurand x)

$$V = f(x)$$

- In the simplest case , *f* is linear
  - Ex. strain gauge:

$$V = k \cdot \varepsilon$$

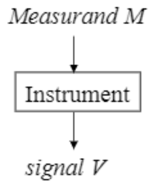
    - V: signal; ε: strain; k: sensor parameter
- The sensor is usable in realiable and efficacy way only when all the sensor parameters are known.
- The estimation of the sensor paramenters can be obtained only with the sensor calibration ..i.e. with a series of experimental measurements and using a statistical regressions.

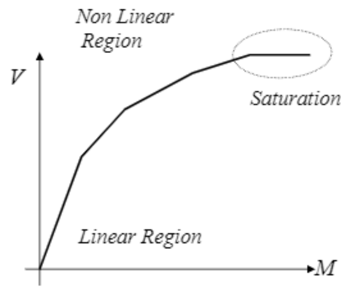
14

14

Fundamental parameters of the instruments: 1. the response curve

- Basically an instrument defines a mapping between the measurand space and output signal space.
- If these spaces have a 1-D, the sensor can be determined through the relationship *V=f(M).*
- This function is defined as the I/O sensor characteristic.
- *The knowledge of this curve allow to use the sensor as measurement tool: from the output signal we can obtain information about the value of x*
- The response curve is obtained with the calibration procedure.

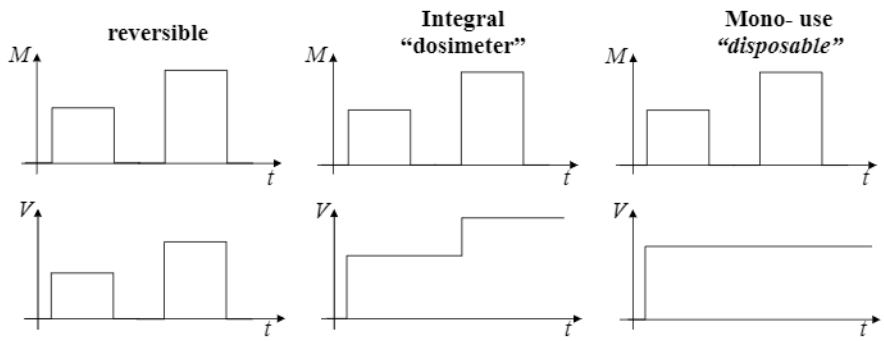




15

Fundamental parameters of the instruments: 2.Reversibility

- The reversibility expresses the capacity of the measuring instrument to follow, with a its typical dynamic, variations of the measurand.
- In particular a instrument is reversible if the response drops to zero when the input ceases .



16

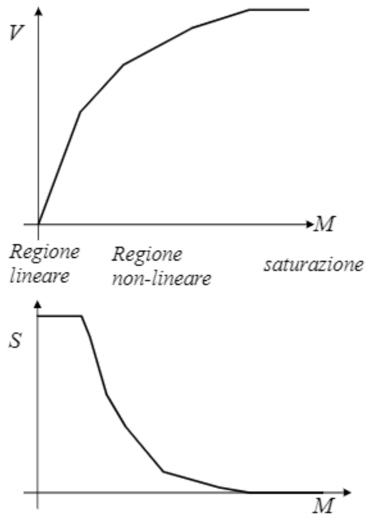
16

Fundamental parameters of the instruments: 3.Sensibility

- Sensibility is defined as the ratio between the signal changes and the measurand
- It defines the capability of the instrument to follow the variation of the measurand
- From mathematical point of view it is the derivate of the response curve of the instrument ,

$$S\big|_{M=M_0} = \frac{dV}{dM}\bigg|_{M=M_0}$$

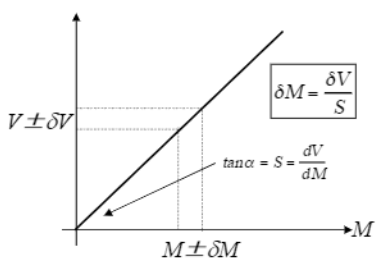
- In the non linear region S is dependent from the working point (i.e. from the measurand value).



17

Fundamental parameters of the instruments: 4. Resolution

- The resolution is correlated with measurement errors and with the noise.
- For this reason, the sensor signal is not a deterministic quantity but it has a random part : V±dV, where dV represents the contribution of all the measurement errors.
- dV is bounded below from the electrical noise of the V.
- The resolution is the relation between the uncertainly of dV and that of dM.



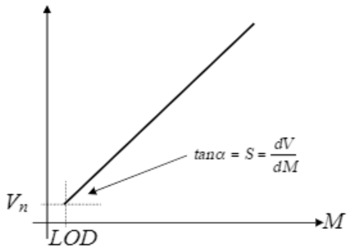
$$\text{Resolution}\big|_{M=M_0} = \lim_{V_{out} \to V_{noise}} \frac{V_{out}}{S\big|_{M=M_0}} = \frac{V_{noise}}{S\big|_{M=M_0}}$$

- The resolution is a function of Sensibility.
- In instruments with high sensibility, the measurement errors are less influences on the measurand estimation.

18

Fundamental parameters of the instruments: 5. Limit of Detection

- The resolution is calculated for a signal equal to zero, defines the *limit of detection* (LOD) of the instrument.
- When the measurement errors reaches the inferior limit , the electronic noise V_n, we can calculate the theoretical LOD.
- It is possible to define $LOD_{conventional}$ = (3 o 9)*LOD.

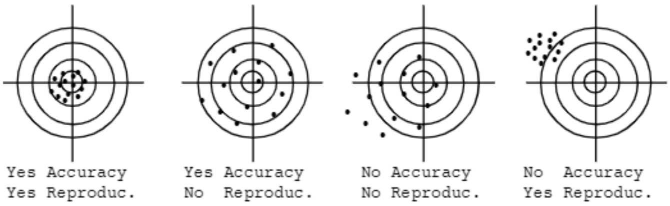


$$LOD = \frac{V_n}{S|_{M \to 0}}$$

19

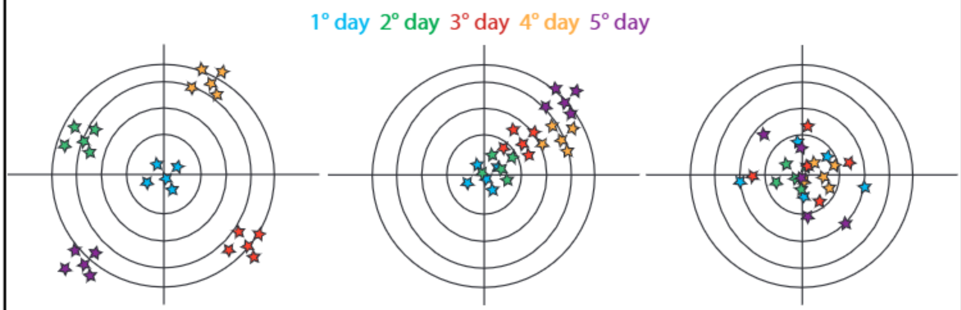
Fundamental parameters of the instruments: 6. Accuracy and Reproducibility

- Accuracy: the capability of a measurement system to give a value of measurand equal to the real value (a priori unknown)
- Reproducibility: the capability of a measurement system to give the same signal in presence of the experimental conditions.
- They are statistical quantities: given N measures the average value is related to the accuracy , the variance of data is related the reproducibility.



20

A step close to real measurements: Accuracy and reproducibility vs the time



Which is the best solution?

21

Fundamental Concepts of Measurement Theory

- **Measurement errors: we obtain different results repeating the "same measure".**
- *For "same measure"… **we are talking about the exposure of the sensor/instrument to the same stimulus (measurand) without the control of the environmental condition that can influence the final response.***
- Response Fluctuations: the mean is the quantity that is the best representation of the measure.
  - Bigger is the number of replicated measurements more reliable is the estimation of the true value of the measurand.

22

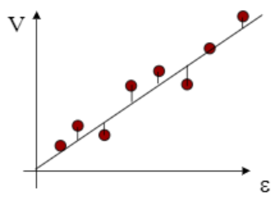
Estimation of the functional parameters

- The characteristic of the functional parameters has to be a-priori known .
- **The deviations from the functional form of the experimental data are considered as measurement errors.**
- To determine the functional form is necessary to do the calibration procedure.
- It consists of applying a sequence of well-known stimuli (often called standards) as input and measuring with high precision the output signal.
- It is straightforward to understand that better is the precision of the standards more accurate will be the estimation of the parameters of the functional form and then of the prediction of the measurand.

23

Estimation of the functional parameters example: strain gauge

Measurand: strain (DL/L)
Output: Voltage
Functional parameter: linear
Parameter: k (gauge factor, sensibility)



$$V = k \cdot \varepsilon$$

calibration — V: known; ε: known; k: unknwon

$$k = \frac{V}{\varepsilon}$$

measure — V: known; ε: unknwon; k: known

$$\varepsilon = \frac{V}{k}$$

Starting from the evidence that there are the measurement errors, it is not possible to apply the formulae on the left but we have to consider a statistical regression that minimizes the error in the estimation of k.

24

Destructive and non destructive analysis

The sample is destroyed or more in general modified and it cannot use for its initial application.

The measurement don't alter the sample in any of its chemo-physical characteristics.

- N.B.: in most cases, the success in industrial application (real scenario) of a specific measurement techniques can be strongly influenced by the destructive or not destructive feature of the measurements.
  - Example : sugar concentration in the water-melon!!

25