2.2 Data analytics for regression and classification: Projection on Latent Structures (PLS)

Prof. Pierantonio Facco
University of Padova, Italy
pierantonio.facco@unipd.it

GRICU PhD School 2021

Digitalization Tools for the Chemical and Process Industries

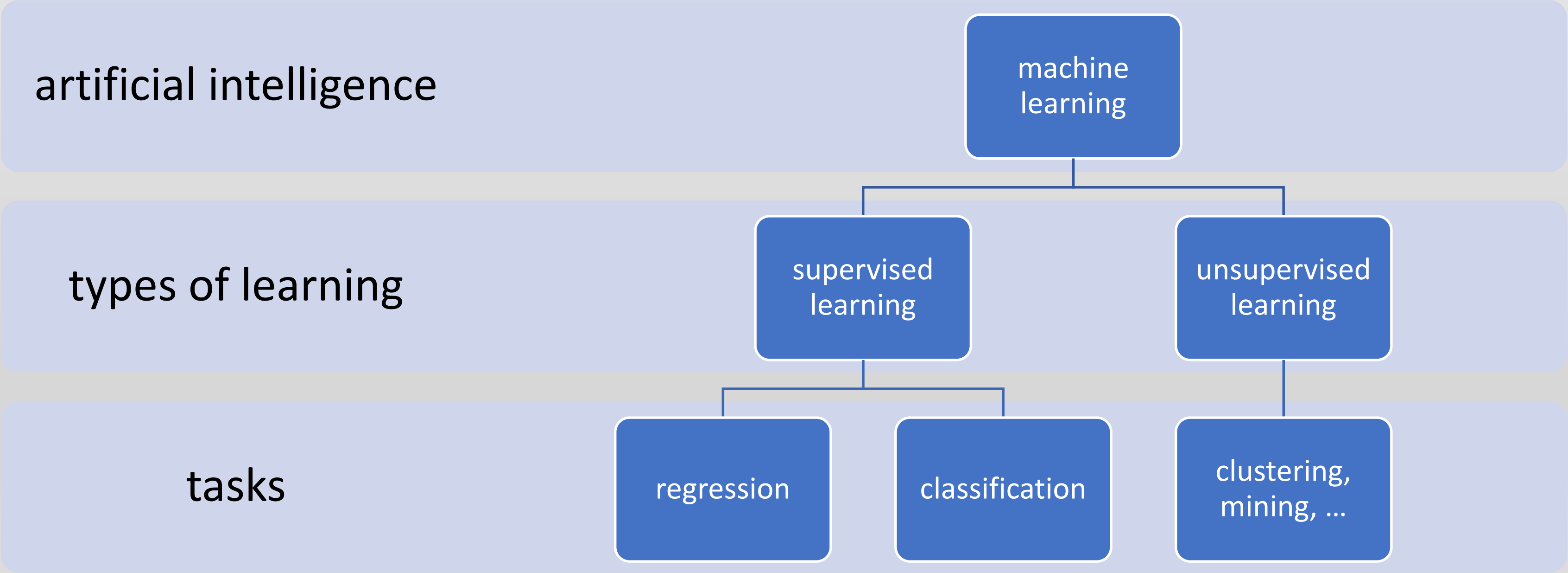
March 12, 2021

Outline

- Contextualization and motivation
- Latent variables methodologies
- Partial Least Squares (PLS)
  - formulation
  - geometrical interpretation
- Example

Machine learning

- **Machine learning** is a branch of artificial intelligence that exploits algorithms and statistical models that computer systems use to effectively perform a specific task without being explicitly programmed to perform that task, but learning from data:
  - **unsupervised learning**: data grouping and interpretation based on one data type
  - **supervised learning**: predictive models based on inputs-outputs relations



Unsupervised learning for exploration and mining

- Data exploration and mining are used for data general **overview and summary**:
  - how the observation are related
  - detection of deviating observations
  - identification of different data classes (clusters)
  - understanding on the relationship between variables
  - assessing if some variables contribute in similar manner on observations
  - understanding similarities and dissimilarities among observations
  - etc…

Supervised learning

- The problem is observing *if a set of input variables (which are measured or preset) have some influence on one or more outputs*:
  - the **inputs X** are often called (the terms can be used interchangeably):
    - predictors
    - independent variables
    - factors (in Design of Experiments)
    - features (in the pattern recognition literature)
    - regressors
  - the **outputs Y** are called:
    - responses
    - dependent variables

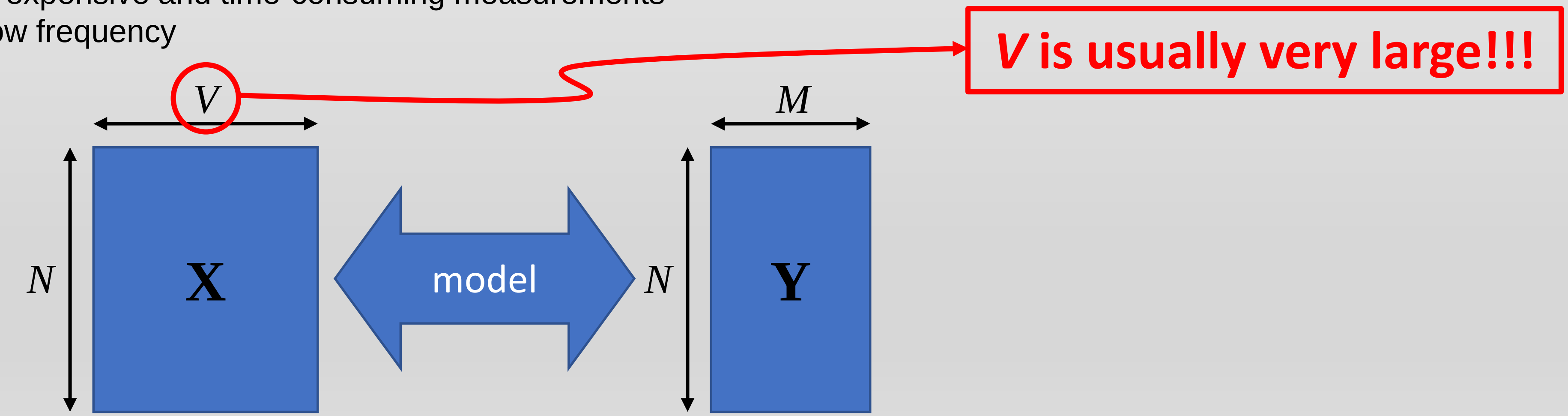
- The **goal** of **supervised learning** is to use the inputs to **predict/estimate** the values of the outputs

Regression and classification

- For all the types of inputs and outputs pairs inputs are used to predict/estimate the output:
  - examples:
    - given specific biological and chemical measurements (e.g.: viability, pH, dissolved oxygen) in the previous days of culture, the titer of an experiment can be predicted
    - given todays' weather conditions, wind intensity and direction and environmental humidity, tomorrow's weather can be forecasted
    - depending on ingredients, water content, powders drying temperature and compaction pressure, the class of crumbliness of a paracetamol tablet can be predicted
- Conventionally, different prediction/estimation tasks (which have a lot in common!) are determined by distinct output types:
  - **regression**: **quantitative outputs** prediction/estimation
  - **classification**: **qualitative outputs** prediction/estimation
    - both can be viewed as a task in function approximation

Estimation and prediction

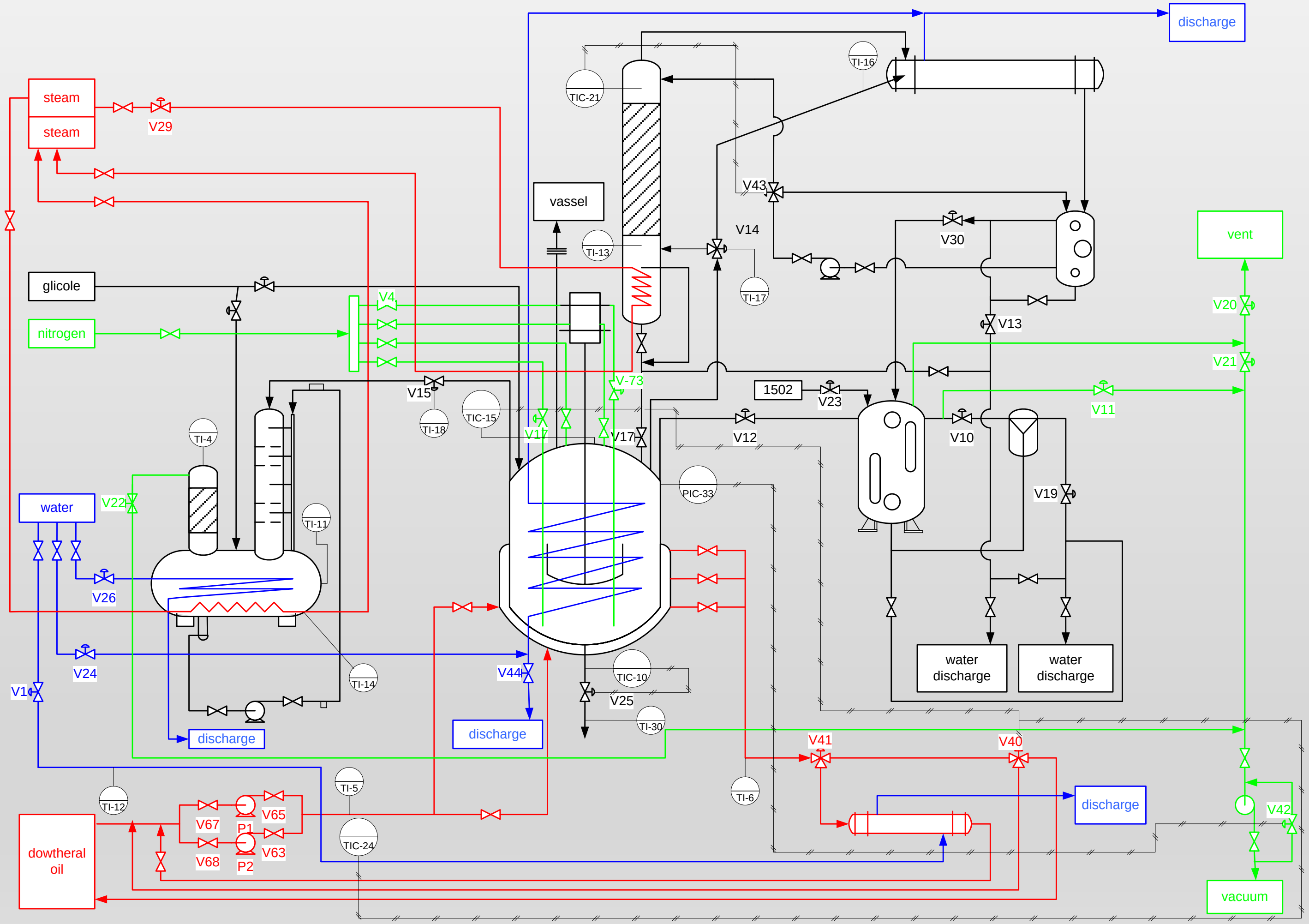
- Estimation and prediction are carried out by means of **classification and regression models**:
  - find out how predictors are quantitatively related to responses
  - give information on how factors can be used to adjust responses
- Two blocks of data are modelled
  - **predictors** (or factors): $\mathbf{X}\ [N \times V]$
    - usually sampled frequently and at regular intervals
  - **responses** that are estimated: $\mathbf{Y}\ [N \times M]$
    - often laborious, expensive and time-consuming measurements
    - available with low frequency



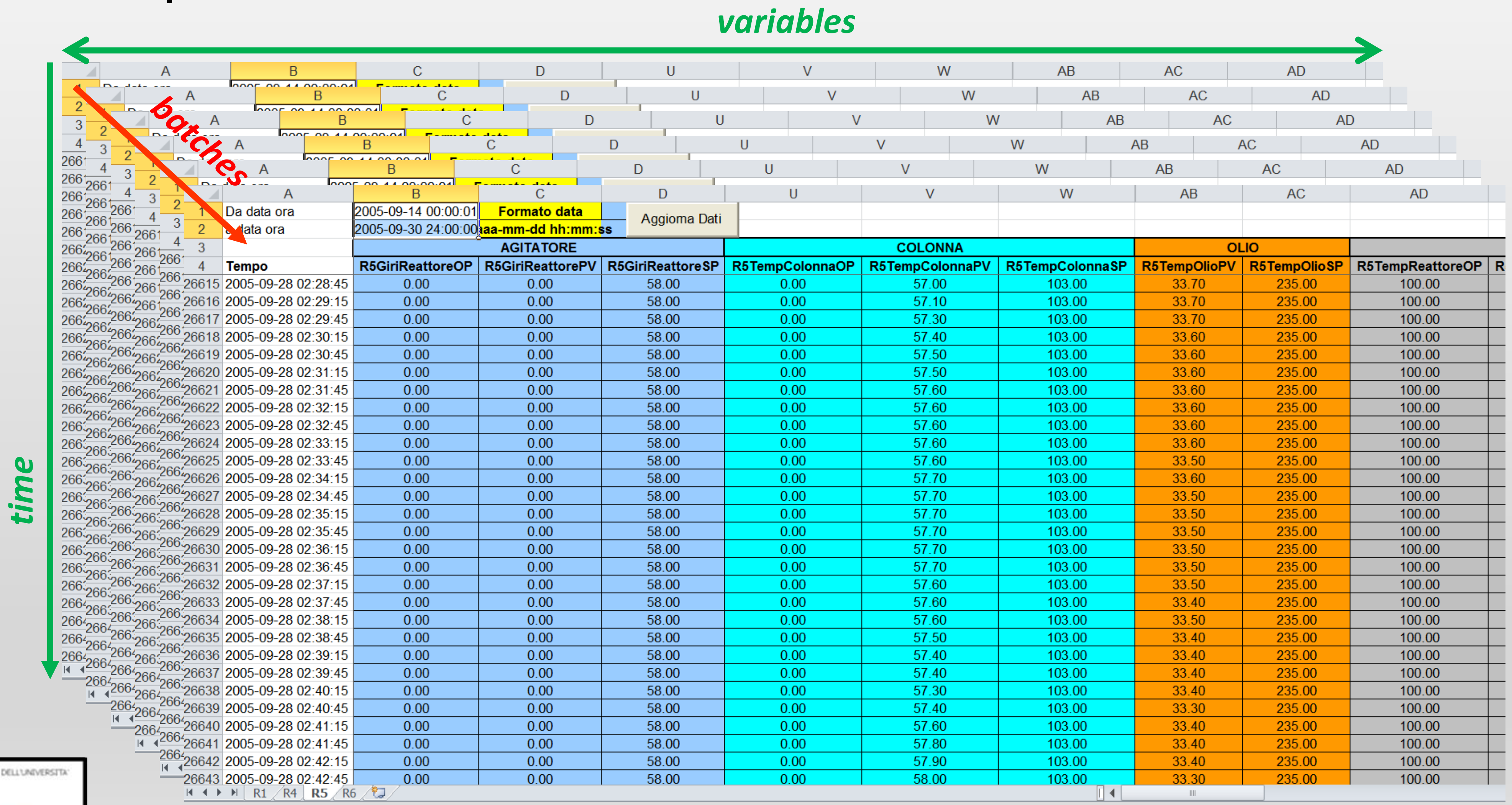
Data challenges

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 $V > 3$
     - 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, qualitative and ordered categorical
     - quantitative may assume any reasonable real value in a continuous scale
     - qualitative are categorical variables that assumed predetermined levels
     - ordered categorical have an ordering between values, but no metric notion is appropriate
   - 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

Chemical plant for resin manufacturing



Online plant instrumentation data



		AGITATORE			COLONNA			OLIO		
	Tempo	R5GiriReattoreOP	R5GiriReattorePV	R5GiriReattoreSP	R5TempColonnaOP	R5TempColonnaPV	R5TempColonnaSP	R5TempOlioPV	R5TempOlioSP	R5TempReattoreOP
26615	2005-09-28 02:28:45	0.00	0.00	58.00	0.00	57.00	103.00	33.70	235.00	100.00
26616	2005-09-28 02:29:15	0.00	0.00	58.00	0.00	57.10	103.00	33.70	235.00	100.00
26617	2005-09-28 02:29:45	0.00	0.00	58.00	0.00	57.30	103.00	33.70	235.00	100.00
26618	2005-09-28 02:30:15	0.00	0.00	58.00	0.00	57.40	103.00	33.60	235.00	100.00
26619	2005-09-28 02:30:45	0.00	0.00	58.00	0.00	57.50	103.00	33.60	235.00	100.00
26620	2005-09-28 02:31:15	0.00	0.00	58.00	0.00	57.50	103.00	33.60	235.00	100.00
26621	2005-09-28 02:31:45	0.00	0.00	58.00	0.00	57.60	103.00	33.60	235.00	100.00
26622	2005-09-28 02:32:15	0.00	0.00	58.00	0.00	57.60	103.00	33.60	235.00	100.00
26623	2005-09-28 02:32:45	0.00	0.00	58.00	0.00	57.60	103.00	33.60	235.00	100.00
26624	2005-09-28 02:33:15	0.00	0.00	58.00	0.00	57.60	103.00	33.60	235.00	100.00
26625	2005-09-28 02:33:45	0.00	0.00	58.00	0.00	57.60	103.00	33.50	235.00	100.00
26626	2005-09-28 02:34:15	0.00	0.00	58.00	0.00	57.70	103.00	33.60	235.00	100.00
26627	2005-09-28 02:34:45	0.00	0.00	58.00	0.00	57.70	103.00	33.50	235.00	100.00
26628	2005-09-28 02:35:15	0.00	0.00	58.00	0.00	57.70	103.00	33.50	235.00	100.00
26629	2005-09-28 02:35:45	0.00	0.00	58.00	0.00	57.70	103.00	33.50	235.00	100.00
26630	2005-09-28 02:36:15	0.00	0.00	58.00	0.00	57.70	103.00	33.50	235.00	100.00
26631	2005-09-28 02:36:45	0.00	0.00	58.00	0.00	57.70	103.00	33.50	235.00	100.00
26632	2005-09-28 02:37:15	0.00	0.00	58.00	0.00	57.60	103.00	33.50	235.00	100.00
26633	2005-09-28 02:37:45	0.00	0.00	58.00	0.00	57.60	103.00	33.40	235.00	100.00
26634	2005-09-28 02:38:15	0.00	0.00	58.00	0.00	57.60	103.00	33.50	235.00	100.00
26635	2005-09-28 02:38:45	0.00	0.00	58.00	0.00	57.50	103.00	33.40	235.00	100.00
26636	2005-09-28 02:39:15	0.00	0.00	58.00	0.00	57.40	103.00	33.40	235.00	100.00
26637	2005-09-28 02:39:45	0.00	0.00	58.00	0.00	57.40	103.00	33.40	235.00	100.00
26638	2005-09-28 02:40:15	0.00	0.00	58.00	0.00	57.30	103.00	33.40	235.00	100.00
26639	2005-09-28 02:40:45	0.00	0.00	58.00	0.00	57.40	103.00	33.30	235.00	100.00
26640	2005-09-28 02:41:15	0.00	0.00	58.00	0.00	57.60	103.00	33.40	235.00	100.00
26641	2005-09-28 02:41:45	0.00	0.00	58.00	0.00	57.80	103.00	33.40	235.00	100.00
26642	2005-09-28 02:42:15	0.00	0.00	58.00	0.00	57.90	103.00	33.40	235.00	100.00
26643	2005-09-28 02:42:45	0.00	0.00	58.00	0.00	58.00	103.00	33.30	235.00	100.00

Laboratory data

variables

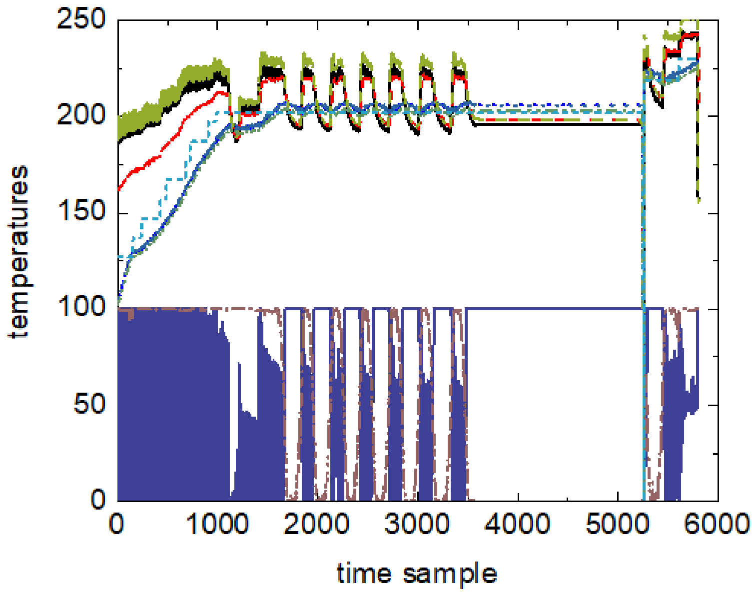
DATA-ORA	m3/h	Reattore teorica	Reattore reale	Colonna SP/PV=120	distillazione su colonna rientro totale			CAP 2000. 100.C, cono 4, 2000.rpm, 25"
28/09/2006 00:00			30		**inizio cottura giovedì 28/09/06**			
28/09/2006 04:00	min. giù		100		caricato materie prime			
28/09/2006 06:45	min. giù	127	127		Inizio distillazione			
28/09/2006 07:45	min. giù	137	131	100.3				
28/09/2006 08:45	min. giù	147	138	100.3				
28/09/2006 09:45	min. giù	157	146	100.8				
28/09/2006 10:45	min. giù	167	153	100.8				
28/09/2006 11:45	min. giù	177	160	101.8				
28/09/2006 12:45	min. giù	187	168	103.1				
28/09/2006 13:45	min. giù	197	173	104				
28/09/2006 14:45	min. giù	202	177	103.4				
28/09/2006 16:30	min. giù	202	185	101.9	Fare prelievo		28.1	1.11
28/09/2006 19:00	min. giù	202	200	99.4			19.3	1.793
28/09/2006 21:00	min. giù	202	202				16	2.295
28/09/2006 21:30	min. giù	202	202		**chiuso colonna, passati in via breve**			
28/09/2006 23:30	min. giù	202	202				14	2.79
29/09/2006 01:30	min. giù	195	202				13.2	3.12
29/09/2006 03:30	min. giù	202	202				11.3	3.398
29/09/2006 03:40	min. giù	202	202		**agg. 6066 + 25Kg. 6058 passati via sruber**			
29/09/2006 06:00	min. giù	202	202				9.59	3.473
29/09/2006 06:30	min. giù	202	202		inserito vuoto			
29/09/2006 08:30	min. giù	202	202				8.08	3.938
29/09/2006 10:30	min. giù	202	202				6.82	4.5
29/09/2006 12:30	min. giù	202	202		CAP RPM 750		5.81	5.36
29/09/2006 13:00	min. giù	202	202		**rotto vuoto Agg 6066 + 3Kg. 6058**			
29/09/2006 15:00	min. giù	202	202		**INSERITO VUOTO**		5.09	5.98
29/09/2006 17:00	min. giù	202	202				4.58	6.4
29/09/2006 20:00	min. giù	202	202				3.95	7.04
29/09/2006 23:00	min. giù	202	202		**Acidità:3,5-3,7Alzato temperatura e alzato Azoto**		3.3	7.66
30/09/2006 02:30	min. giù	218	218				3	8.44
30/09/2006 05:30	min. giù	218	218				2.4	9.22
30/09/2006 06:30	min. giù	230	218		**Acidità: <2,3 Alzato temperatura**		2.3	9.46
30/09/2006 08:00	min. giù	230	218				2.24	9.76
30/09/2006 09:30	0,3 VR	230	219				2.13	10.24
30/09/2006 10:00	0,3 VR	230	219				2.09	10.36
30/09/2006 10:15	0,3 VR	230	220		rotto vuoto aperto raffreddamento		1.99	10.42

time

observations

Online variables in the production of resins

- 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

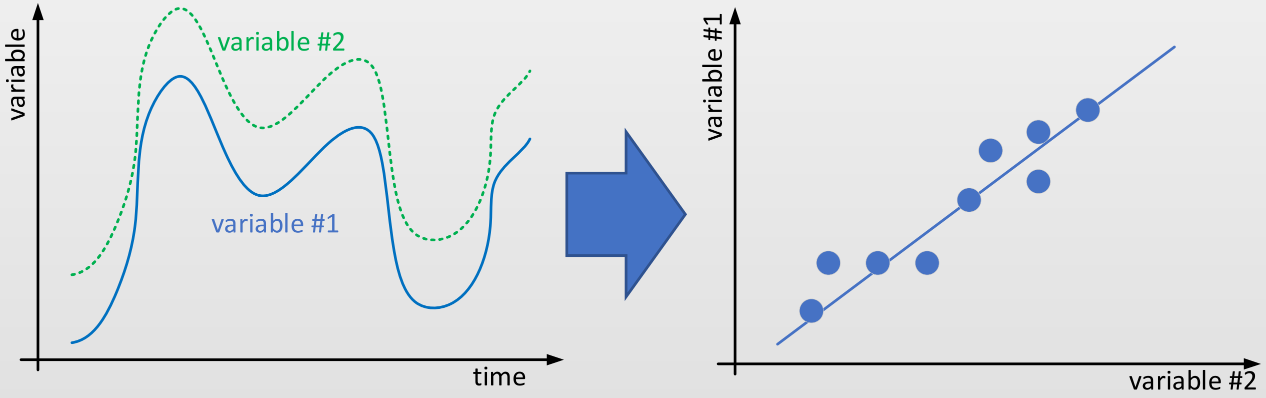
- From the mathematical point of view, the **correlation** among two variables *x* and *y* are is:

$$\rho_{x,z} = \frac{\sigma_{x,z}}{\sigma_x \sigma_z} \in [-1,1] \qquad \begin{cases} \sigma_{x,z} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_x)(y_n - \mu_z) & \text{covariance} \\ \sigma_x = \frac{1}{\sqrt{N}}\sqrt{\sum_{n=1}^{N}(x_n - \mu_x)^2} \\ \sigma_z = \frac{1}{\sqrt{N}}\sqrt{\sum_{n=1}^{N}(y_n - \mu_z)^2} \end{cases}$$

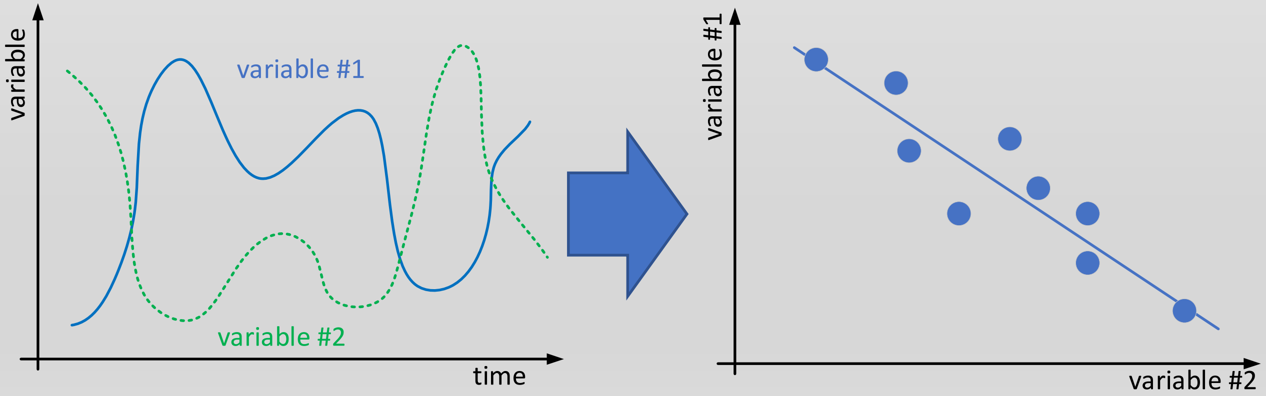
- Evaluating the **correlation structure** in a dataset means observing if data:
  - vary one in strict relation with the others
  - show common behavior (i.e.: trends, shape, etc…)

Correlation in practice

- **High positive correlation** $(\rho \sim 1)$
  - 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



- **High negative correlation** $(\rho \sim -1)$
  - two variables are negatively correlated when they vary in opposite sides
    - when one goes up, the other goes down and vice-versa



GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERSITA'
GRICU
GET INNOVATION

What misses in the univariate thinking?

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

Multivariate data challenges

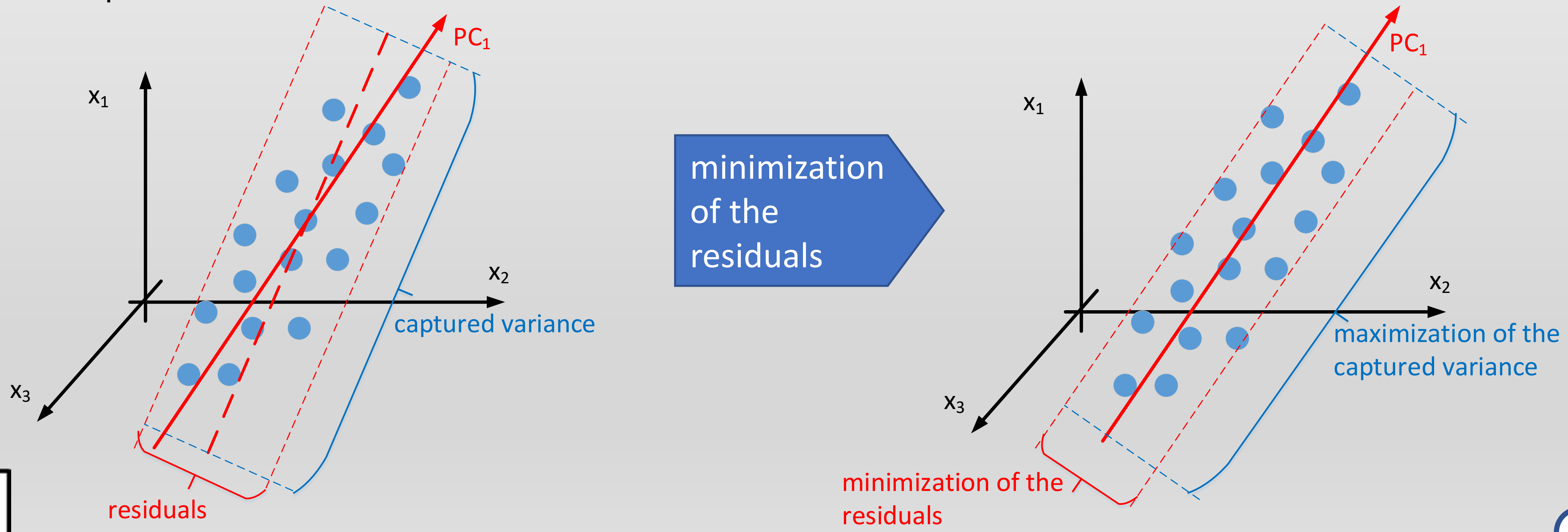
1. **Dimensionality**:
   - **thousands of variables are recorded every few seconds** thanks to digitalization in Industry 4.0
   - all data points are needed for a **proper inspection**
     - do not discard variables or samples if there is not a strong motivation!
     - do not refer to few “reference” variables
2. **Multi-collinearity**:
   - variables are usually **correlated** one another (not straightforward to interpret)
     - **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**
3. **Noise**: unwanted (known or unknown) variability
   - important effects may be partially obscured by noise
4. **Missing data**:
   - data tables may be partially incomplete (i.e., sensor failures, transducer problems, etc…)

Projection-based latent variables models

- Allow **dimensionality reduction**:
  - **compress the data dimensionality** from the original space of *V* variables to a much reduced **space of $A << V$ latent variables LVs**
  - LVs represent the **physical phenomena affecting the system/ process**
- Identify the **correlation between variables**:
  - use variables correlation to compress original variables in latent variables
- Identify the **direction of maximum variability of the data**:
  - approximate the data through an **optimal fitting** (i.e., high representativeness of the model)
- **Filter noise**:
  - discard the non-systematic part of the signals
- **Handle missing data**

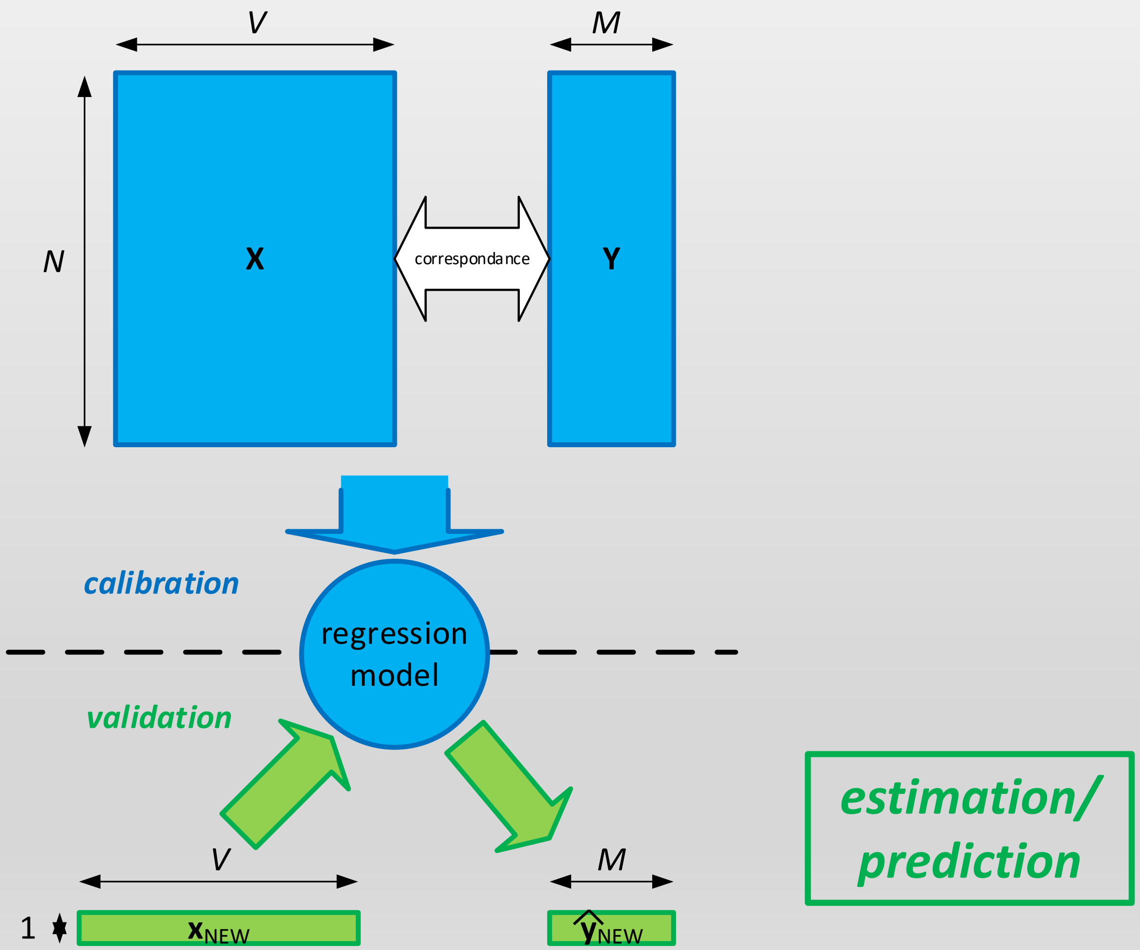
Latent variables models (LVM) ontology

- Latent variables best fit the data points in the space of the original variables
  - find the lines/planes/hyperplanes that best approximate the data in the **least-squares sense**
  - **minimize the residuals** of the fitting space
    - this implies the **maximization of the coordinates variance**



Multivariate regression models

Schematic of regression/classification problems



Regression problem formalization

- **Inputs data X** (i.e., regressors, predictors, independent variables) are available (with high frequency)
- The **corresponding output data y** (i.e., regressed variables, estimated/predicted variables, dependent variables for the same observed units) are available, as well

REGRESSION MODEL $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$

- Regression parameters have to be estimated from the available data
- The **estimated regression parameters** $\hat{\boldsymbol{\beta}}$ are used to **estimate/predict the response variable** $\hat{y}_{\mathrm{NEW}}$ for new observations whose predictors $\mathbf{x}_{\mathrm{NEW}}$ are available

GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERSITA' GRICU GET INNOVATION

Least-squares solution for parameter estimation

- The estimated regression coefficients in *vector form* are:

$$RSS = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\left.\frac{\partial RSS}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

when correlated variables are present in matrix X this equation cannot be solved!

- The fitted regression model is:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

  - and for the *n*-th observation is:

$$\hat{y}_n = \mathbf{x}_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

Projection on Latent Structures, PLS

Partial Least Squares

Partial Least Squares PLS

- PLS (a.k.a. Projection on Latent Structures) is a **linear regression technique** for the association between **X** and **Y**
  - exploits the typical ability of the multivariate methods to analyze many **noisy and collinear data**, dealing with the *ill-conditioned regression problems*

- Geometrically, PLS finds lines/planes/hyperplanes of the closest fit for a system of points in the space of **X** that are most related and predictive for the space of **Y**

Mathematical formulation of PLS (1/2)

- PLS is a method which explains the **directions of maximum variability of X that best predict Y**
  - PLS reduces the dimension of the system, simultaneously finding the space of LVs that are more predictive for the secondary variables and are near to the direction of maximum variability of the primary variables
- The method consists of the following relations:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a^{\mathrm{T}} + \mathbf{F}$$

$$\mathbf{u}_a = b_a \, \mathbf{t}_a$$

- **T** and **U** **scores**
- **P** and **Q** **loadings**
- **E** and **F** **residuals**
  - minimized in the least squares sense
- b_a are the **regression coefficients**:

$$b_a = \frac{\mathbf{u}_a^{\mathrm{T}} \mathbf{t}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}$$

- A smart and fast manner to calculate recursively the PLS parameters is the NIPALS algorithm (Geladi & Kowalski, 1986, Anal. Chim. Acta)

Mathematical formulation of PLS (2/2)

- PLS finds a *transformation of the* **X** *data in order to maximize the covariance of its latent variables (LVs) with the* **Y** *dataset variables*
- **For the first LV** this is represented by the following optimization problem:

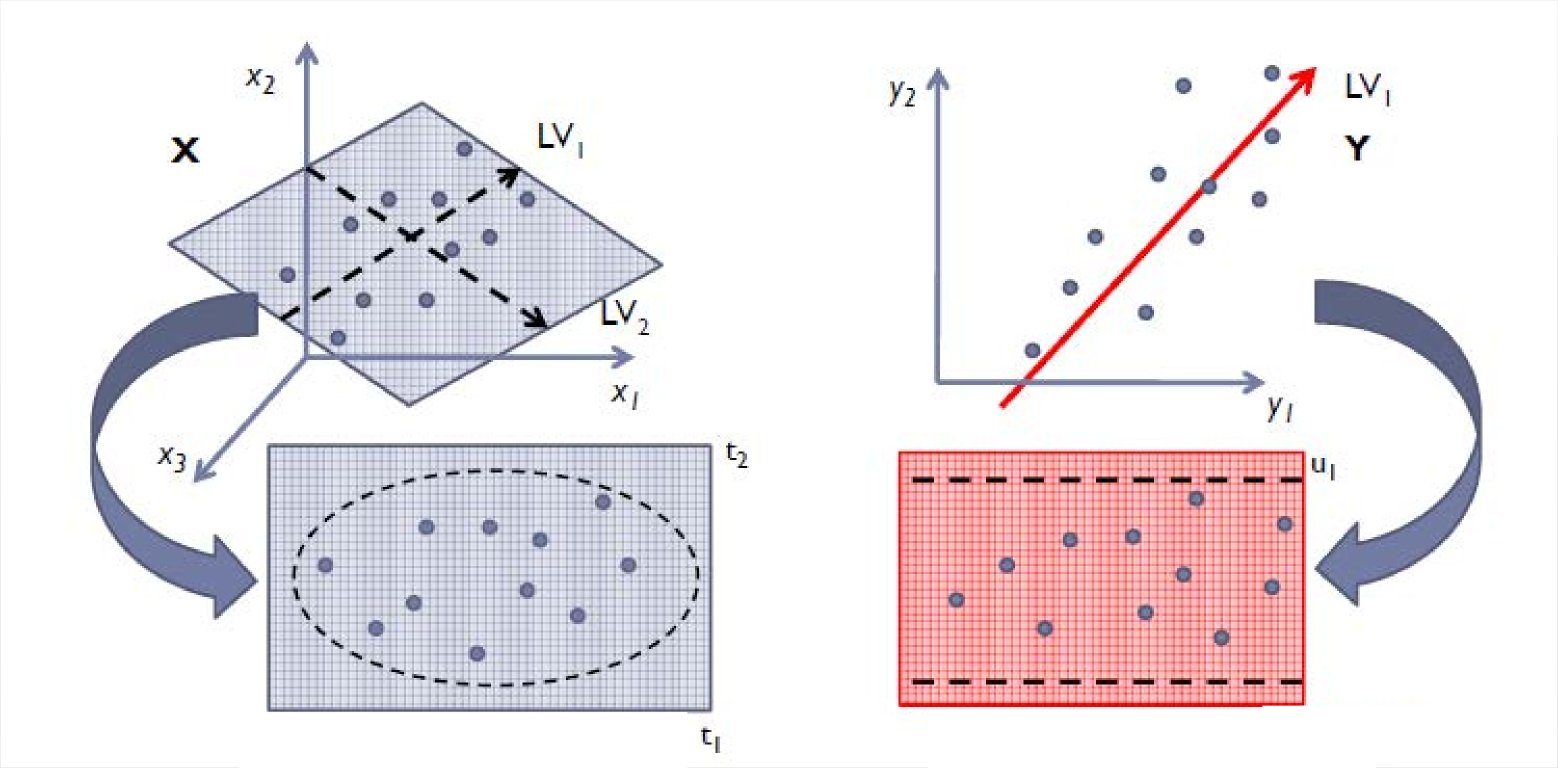
$$\max_{\mathbf{w}_1^*}\left(\mathbf{w}_1^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1^*\right)$$
$$s.t. \quad \mathbf{w}_1^{*\mathrm{T}}\mathbf{w}_1^* = 1$$

  - it maximizes the covariance of the data projections
  - $\mathbf{w}_1$ is the [*M*x1] **weight vector** and represents the coefficient of the linear combination of **X** determining the scores:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1^*$$

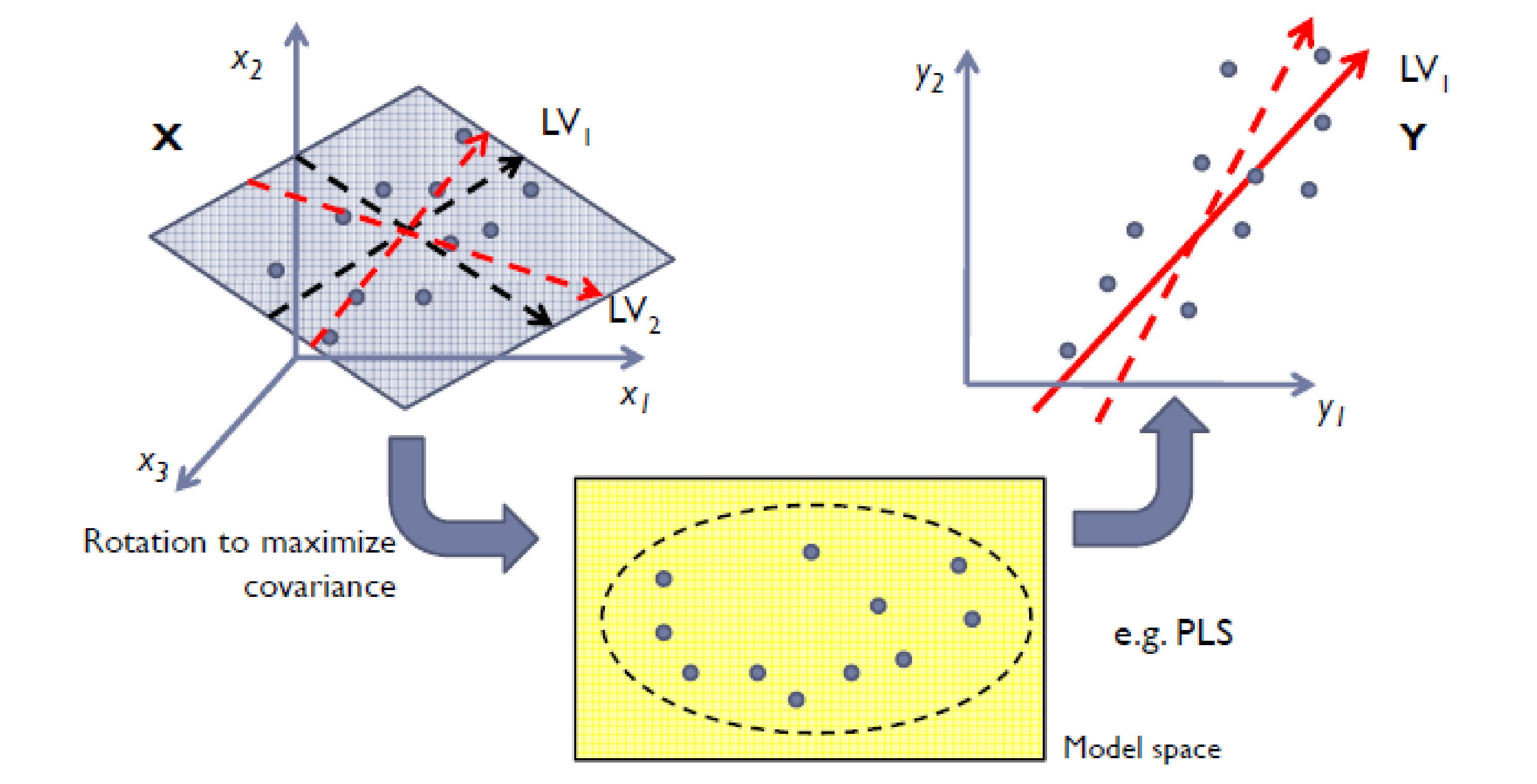
Geometrical interpretation of PLS (1/2)

- Not only PLS finds the direction of maximum variability into the **X** data…



Geometrical interpretation of PLS (2/2)

- … but also rotates them to *optimally predict* **Y**



Interpretation of PLS

- **T** and **U** scores: projections of the observations in the space of the latent variables (i.e., the coordinates in the LV space)
  - **identify the relation among observations**
- **P** and **Q** loadings: are the LVs director cosines
  - **identify the correlation between variables**

} subject to the correlation structure among **X** and **Y**

- **E** and **F** residuals: represent the fitting error
  - minimized in the least-square sense
  - define the distance out of the model hyperspace (i.e., the correlation structure outside the LV space)

GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERSITA' GRICU GET INNOVATION

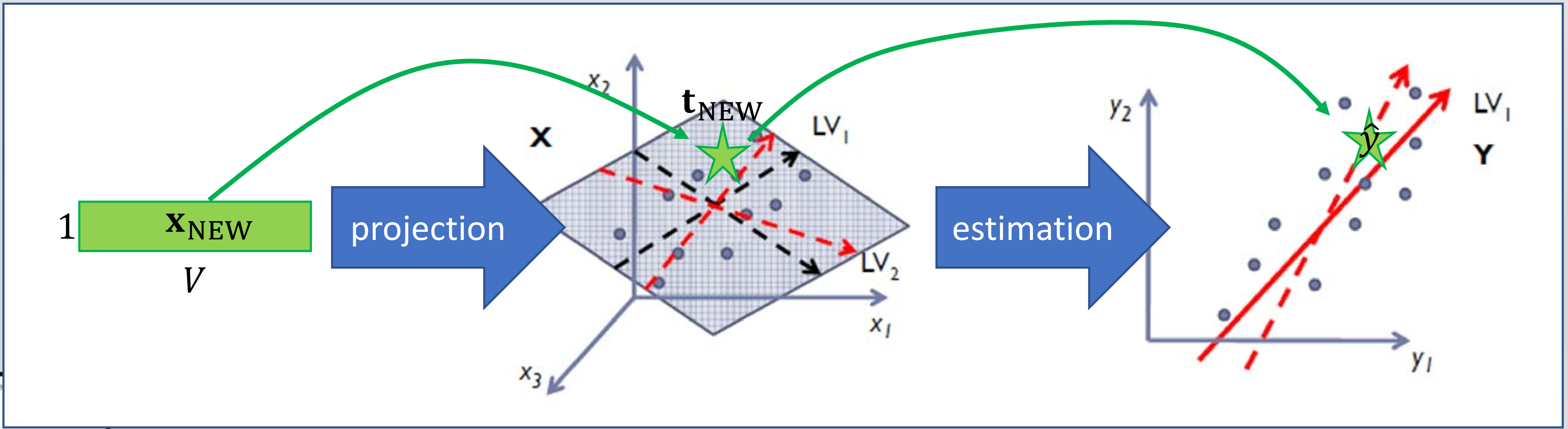
Estimations and predictions

- When a new set of predictors $\mathbf{x}_{\mathrm{NEW}}$ is available it is possible to project it into the space of the LVs:

$$\mathbf{t}_{\mathrm{NEW}} = \mathbf{x}_{\mathrm{NEW}}\mathbf{P}$$

- Then it is possible to predict/estimate the response through:

$$\hat{y} = \mathbf{b}\mathbf{t}_{\mathrm{NEW}}\mathbf{Q}^{\mathrm{T}}$$



PLS diagnostics

- Is a PLS model appropriate to represent the original data with few LVs? How do LVs fit the original data?
- How do observations conform to the correlation structure of the other data in the model?

- **Model diagnostics**
- **Sample diagnostics**

Model diagnostics

- **Model diagnostics**
  - **coefficient of determination**: the amount of variability of the original data explained by the model (in calibration)

$$R^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{n=1}^{N}\sum_{v=1}^{V}\left(x_{n,v} - \hat{x}_{n,v}\right)^2}{\sum_{n=1}^{N}\sum_{v=1}^{V}\left(x_{n,v} - \bar{x}_{v}\right)^2}$$

    - where $\hat{x}_{n,v}$ is the $[n, v]$ element of $\hat{\mathbf{X}}$ and $\bar{x}_v$ is the mean of variable *v*
    - computed for both **X** and **Y**
  - the Q^2 **index**: a measure of the **predictive power** of the model on new unknown samples

$$Q^2 = 1 - \frac{PRESS}{TSS}$$

  - usually R^2 increases with the number of PCs included into the model, while $Q^2 < R^2$ reaches a maximum with the optimal number *A* of LVs

GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERSITA' GRICU GET INNOVATION

Sample diagnostics

- **Sample diagnostics**: computed for both **X** and **Y**:
  - the **Hotelling's *T*² statistic** measures the overall distance of the projections of an observation from the LV space origin (i.e., similarity to the average)
    - LVs explain different data variance aliquots ➡ the Mahalanobis distance is used:

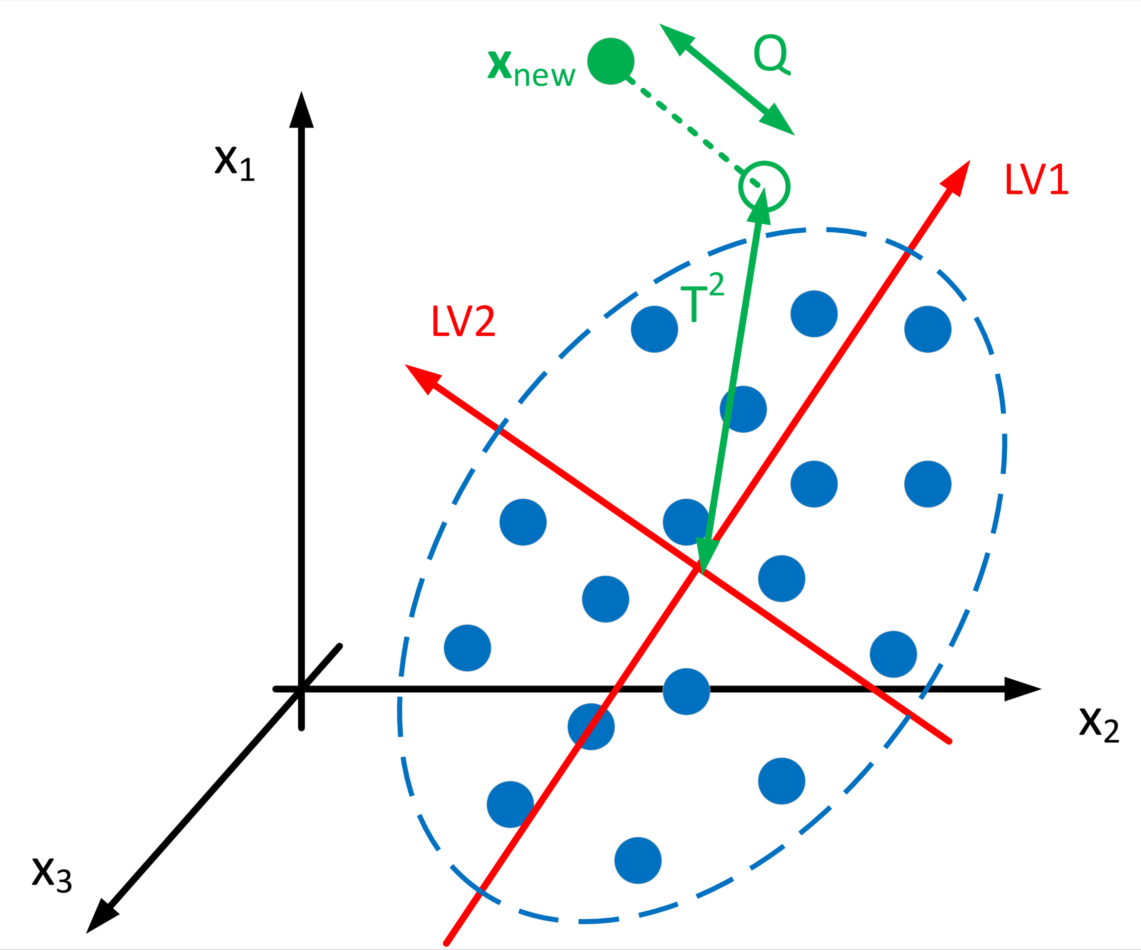
$$T_n^2 = \mathbf{t}_n^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \mathbf{t}_n = \sum_{a=1}^{A} \frac{t_{a,n}^2}{\lambda_a}$$

      - where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues
  - the **squared prediction error Q** (or SPE) measures the orthogonal distance of the n^{th} observation from the latent space of the model
    - measures the representativeness of the model for the observation

$$Q_n = \mathbf{e}_n^{\mathrm{T}} \mathbf{e}_n$$

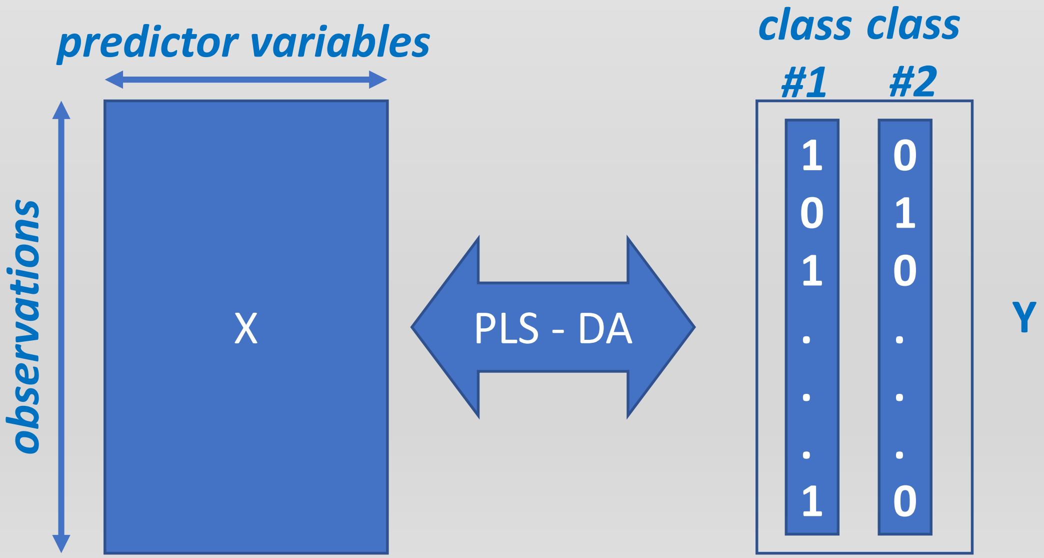
GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERSITA' GRICU GET INNOVATION

Geometrical interpretation of sample diagnostics



PLS-Discriminant analysis

- **PLS- Discriminant Analysis** is the PLS version for **classification**:
  - the response variable expresses the class
    - often with a dummy variable (0-1)
  - the weights identify how the classes are determined by the predictors pattern
  - a probabilistic attribution to the class is possible



PLS common applications

- PLS is commonly applied to different fields:
  - **soft sensing**
  - **process monitoring**
  - **design and transfer of processes and products** between different scales and production sites
  - **process and product optimization**
  - **DoE and response surface modelling**
  - **QSAR** (Quantitative Structure-Activity Relationship modelling)
  - **instrumentation calibration** (e.g.: Near-Infrared Spectroscopy)

Use of PLS

exploratory	predictive	prescriptive
understanding correlation within datasets	response estimation	design of experiments
understanding relations between datasets	class attribution in PLS Discriminant Analysis	PLS model inversion

GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERS
GRICU
GET INNOVATION

Predictive analysis through PLS

exploratory	predictive	prescriptive
understanding correlation within datasets	**response estimation**	design of experiments
understanding relations between datasets	class attribution in PLS Discriminant Analysis	PLS model inversion

GRUPPO DI INGEGNERIA CHIMICA DELL'UNIVERS

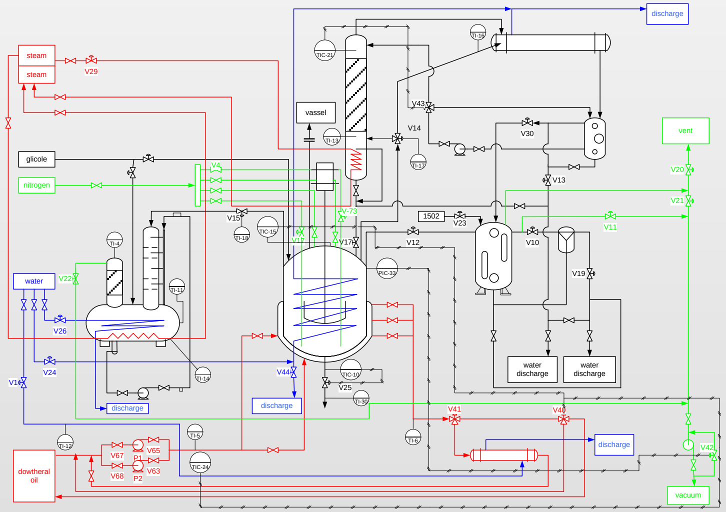
GRICU

GET INNOVATION

Soft sensor in the production of resins for coatings

Batch production of resins for coatings

- Production of **resins for coatings:**
  - semi-batch polymerization
  - 12 m^3 reactors
- Resins' quality indices:
  - **acidity**
  - **viscosity**

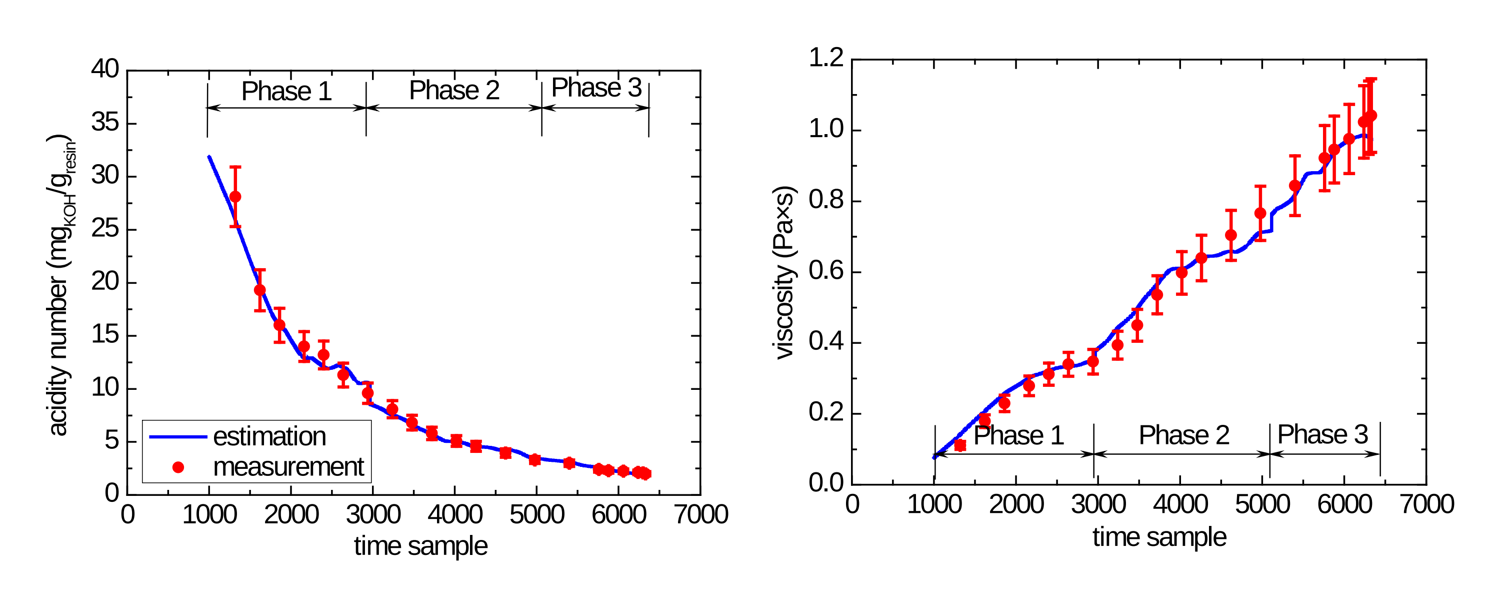


- **Issues**:
  - raw materials variability
  - plant operations are carried out manually from operating personnel
  - manual measurements of quality indices
    - every 2 h
  - manual corrections of the recipe
    - based on operators experience
  - high variability of the batch duration: 40-70 h

- **Objectives**:
  - **real time estimation of the product quality**
    - prompt corrections of the recipe
    - avoid out-of-spec
  - **batch duration prediction** from the initial part of the batch
    - production organization
    - labor resources management

Virtual sensor

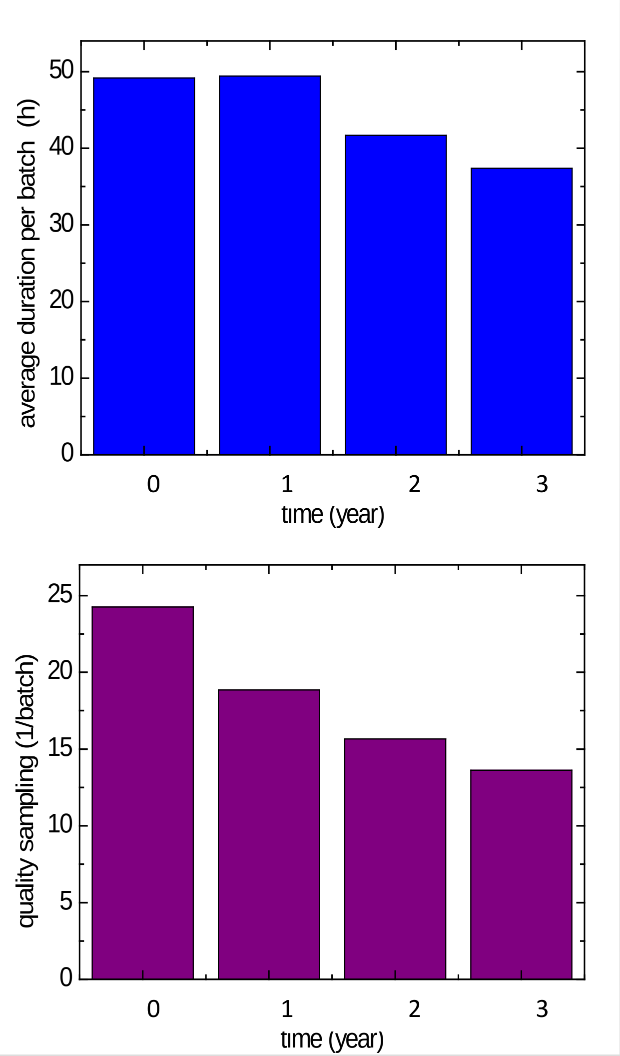
- **Product quality online estimation** (acidity e viscosity)
  - **frequency** every 30 s >> frequency of lab assays
  - **accuracy** = accuracy of lab measurements
- **Batch duration prediction**: accuracy < 4 h already 10 h from the batch start
  - correlated to initial temperature ramps management



Operating and economic benefits

- “Faster” batches: -10 h duration
- Lower number of lab assays: -10 samples/batch
- Total savings:
  - *1000+ lab measurements*
  - *100+ h process*

- **Increased production: 250 000 kg/year**
- **Saving: 340 h/operators**
- + materials, instruments, etc…



Final remarks

- PLS is a powerful tool to correlate two blocks of multivariate data
- PLS can be used to perform accurate:
  - **process understanding** and **correlative analysis**
  - **estimations/predictions** and **classification**
  - **product formulation and process design and scale-up**
- PLS is a **flexible technique** to deal with:
  - hardware sensor data
  - chemical, physical, mechanical measurements
  - panel judgement
  - images
  - spectra
  - internet data, etc…

References

- **Geladi, Kowalski (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 185, 1–17**.
- Wise, Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control* **6**, 329–348.
- Höskuldsson (1988). PLS regression methods. *J. Chemom.* **2**, 211–228.

- **Eriksson, Kettaneh-Wold, Trygg, Wikström, Wold (2006). *Multi- and Megavariate Data Analysis: Part I: Basic Principles and Applications.* Umetrics Inc.**
- Hastie, Tibshirani, Friedman (2008). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction,* 2nd Ed. Springer
- Johnson, Wichern (2007). *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall
- Chiang, Russell, Braatz (2001). *Fault detection and diagnosis in Industrial systems.* Springer