Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #6 – Flipped lecture

Prof. Pierantonio Facco

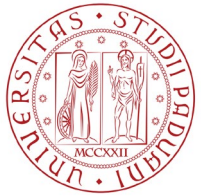
CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Flipped learning procedure

- Today's lecture will be a **flipped lesson**
  - the learning procedure is inverted
- Please, complete the following procedure before Lesson #7:
  1. **attend the video lesson #6** 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** on anything you need to discuss with the teacher and your mates in the following lecture
- The 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



UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF INDUSTRIAL ENGINEERING

Machine Learning Lessons #6 – Part 1

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab

Principal Component Analysis, PCA

What is principal component analysis PCA?

General definition	Mathematical formulation	Model structure	Model and sample diagnostics
• PCA goal • PCA ontology • PCA tasks	• Optimization framework • Singular value decomposition • Iterative calculation procedure • Easy geometrical interpretation	• PCA model structure • PCA model interpretation • Data pretreatment	• Is the model appropriate? • Is an observation conforming to the ones included into the model?

Principal Component Analysis PCA

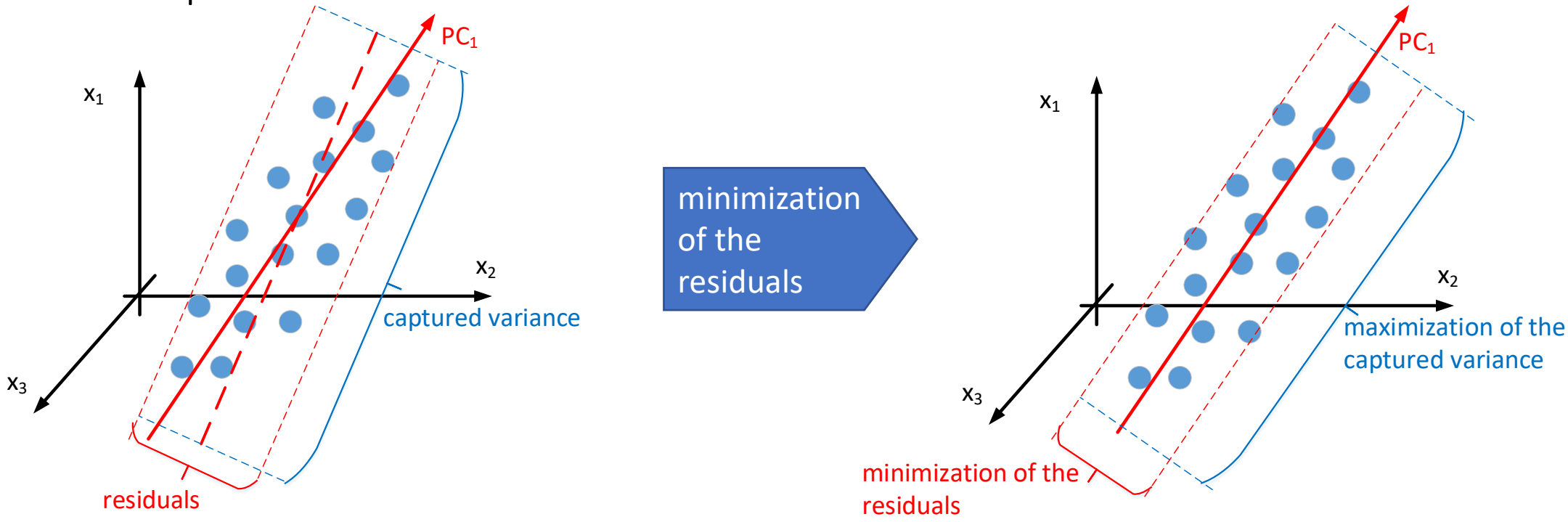
- Principal Component Analysis, **PCA** is the basis of multivariate statistical analysis:
  - deals with one matrix $\mathbf{X}$ $[N\ observations \times V\ variables]$
    - observations can be:
      - analytical samples
      - chemical compounds
      - continuous process time points
      - batches
      - biological individuals
      - DoE trials, etc…
    - variables can be:
      - process variables
      - wavelengths of spectra
      - detection intensity in time of a chromatograms, etc.
- PCA **finds the direction of maximum variability of the data**
  - these directions, called also **principal components** or **latent variables**, describe the hidden information in subsets of variables which are intimately correlated and **describe the main driving forces** of the system under study

Goal of PCA

- The primary goal of PCA is finding lines, planes, or hyperplanes of the closest fit for a system of points in a multidimensional space
  - PCA represents the original observations in a **low-dimensional space** (typically 2-5 latent variables) to obtain a **convenient overview of the data**
    - reveals:
      - groups of observations
      - trends
      - outliers
    - uncovers the relation among:
      - observations
      - variables
      - observations and variables

PCA ontology

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



Mathematical formulation of PCA (1/4)

- PCA (Jackson, 1991):
  - allows **summarizing** the information embedded in a dataset $\mathbf{X}$ $[N \times V]$ of correlated variables
  - projects the data through a **linear transformation onto a new coordinate system** of latent orthogonal variables, whose directions are identified by the **loadings** $\mathbf{P}$:
    - the latent variables **optimally capture the variability of the data and the correlation among the original variables**
  - each of these coordinates identifies a latent direction in the data and is called **principal component PC**, the direction of maximum variance of the data
- To find the directions of the new coordinate system, a combination of the original variables in $\mathbf{X}$ is found which, for the first PC satisfies:

$$\max_{\mathbf{p}_1} \left(\mathbf{p}_1^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{p}_1\right) \quad \textit{optimality condition}$$

$$s.t. \quad \mathbf{p}_1^{\mathrm{T}} \mathbf{p}_1 = 1 \quad \textit{orthonormality condition}$$

  - it maximizes the covariance of the data projections
  - $\mathbf{p}_1$ is the $[V \times 1]$ vector of the combination coefficients, called **loadings**
    - loadings are **linear combinations of the original variables**
  - the loadings are the **director cosines of the PCs**

Mathematical formulation of PCA (2/4)

- The original data can be projected onto the space of the latent variables identified by the loadings, the PCs directions:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$$

  - the **score** vector $\mathbf{t}_1$ $[N \times 1]$ represents the coordinates of the data in the new system of PCs coordinates
  - the PCA problem can be reformulated as:

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

Mathematical formulation of PCA (3/4)

- The analytical solution of the abovementioned maximization problem is the same as the solution of the **eigenvector problem**:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_1 = \lambda_1\mathbf{p}_1$$

  - the loading $\mathbf{p}_1$ corresponds to the **eigenvector** of the covariance/correlation matrix of $\mathbf{X}$ if it is autoscaled*
    - $\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{N-1}$ is the **correlation/covariance matrix** when $\mathbf{X}$ is **autoscaled**
  - λ_1 is the **eigenvalue** associated to the eigenvector $\mathbf{p}_1$
    - λ_1 is an indirect measure of the **variance explained** by the product $\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}$, namely the amount of information embedded in the model by the calculated PC

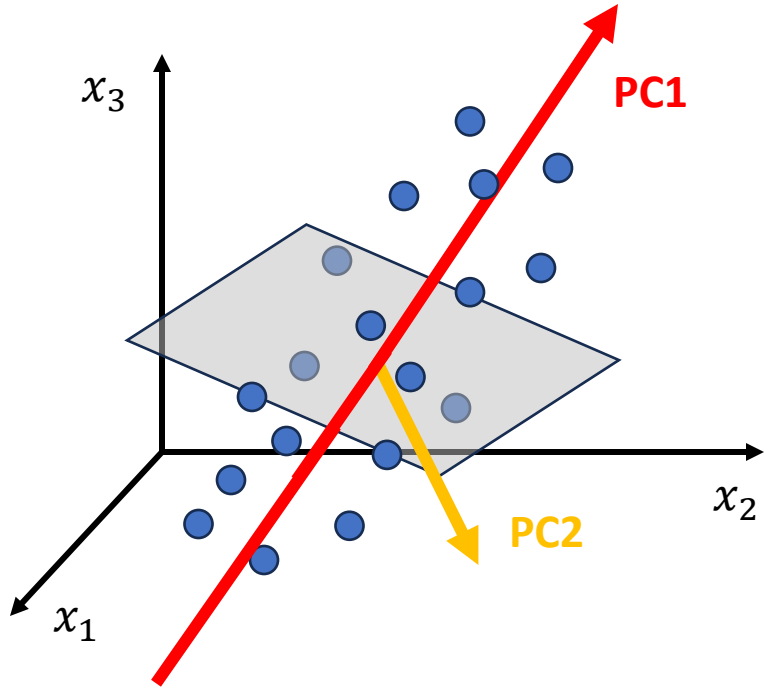
*__autoscaling__ = data normalization: data are **centered to mean zero and scaled to unit variance**; namely, the mean is subtracted from each variable and the centered variable divided by the standard deviation

Mathematical formulation of PCA (4/4)

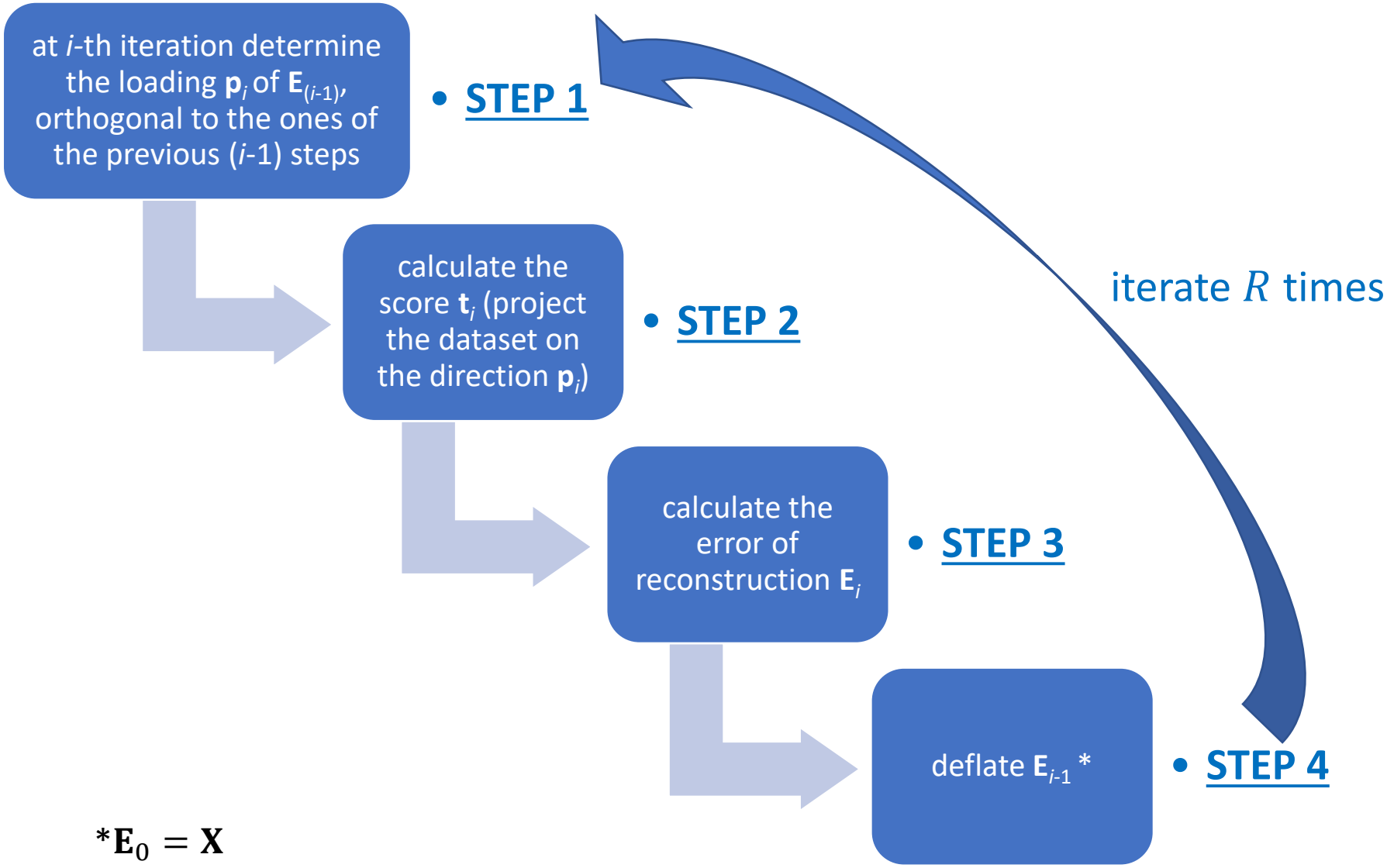
- We can iterate this decomposition to find different principal components
  - at each iteration the original matrix $\mathbf{X}$ is deflated by the reconstruction with the PC previously obtained:

$$\mathbf{E}_1 = \mathbf{X} - \hat{\mathbf{X}}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}$$

  - this determines that different PCs are orthogonal
- The matrix $\mathbf{E}_1$ is the residual matrix and accounts for the reconstruction error of $\mathbf{X}$ (the error made when approximating $\mathbf{X}$ with $\hat{\mathbf{X}}_1$) made by the selected PCs



Principal Components extraction procedure



*$\mathbf{E}_0 = \mathbf{X}$

Summary on the mathematical formulation of PCA

- PCA allows to describe the original dataset $\mathbf{X}$ with a low-dimensional latent space of PCs
  - since the variables of $\mathbf{X}$ are correlated, the $\mathbf{X}$ matrix is not really full rank
  - $\mathbf{X}$ can be represented with a number of PCs $A \ll \min(N, V)$:
    - two or more correlated variables identify a common direction of variability
    - a single PC will therefore capture the variability of a lot of the original variables
    - a single PC is represented by the outer product of scores and loadings

$$\mathbf{X} = \sum_{a=1}^{R} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}}$$

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

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

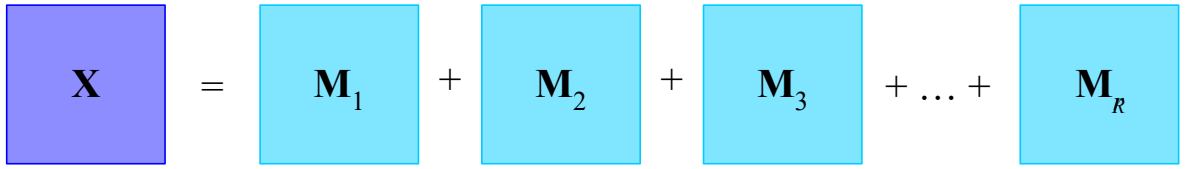
Non-linear Iterative PArtial Least Squares algorithm

- NIPALS algorithm for the a^{th} iteration:
  1. let $\mathbf{x}_v$ be a column of $\mathbf{X}$ and set: $\mathbf{t}_a = \mathbf{x}_v$
  2. calculate: $\mathbf{p}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}}\mathbf{X}}{\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a}$
  3. normalize the loading to unit length: $\mathbf{p}_a^{\mathrm{T}} = \frac{\mathbf{p}_a^{\mathrm{T}}}{\|\mathbf{p}_a^{\mathrm{T}}\|}$
  4. calculate the score: $\mathbf{t}_a = \frac{\mathbf{X}\mathbf{p}_a}{\mathbf{p}_a^{\mathrm{T}}\mathbf{p}_a}$
  5. compare $\mathbf{t}_a$ calculated in 2 to the one calculated in 4:
     - if they are equal for less then an assigned tolerance then the method is converged, else restart from 2 with the last calculated value of $\mathbf{t}_a$
  6. calculate: $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$ with: $\mathbf{E}_0 = \mathbf{X}$
  7. the residual at iteration *a* is the matrix to begin the iteration $(a+1)$, setting:

$$\mathbf{X}_{a+1} = \mathbf{E}_a$$

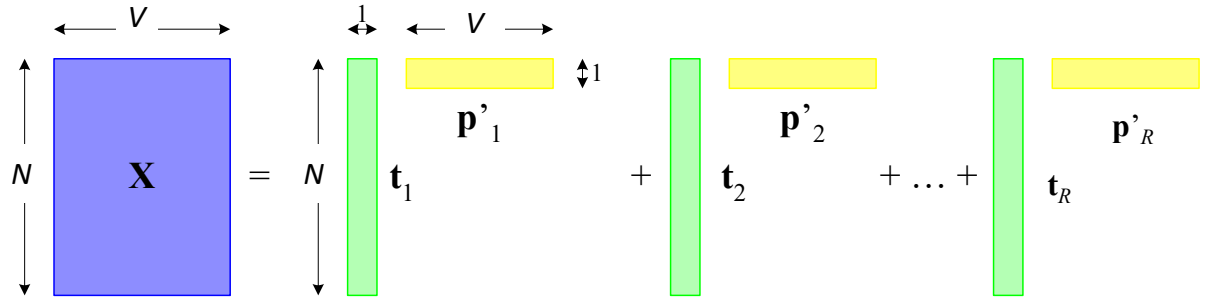
Representation of the PCA formalization

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \cdots + \mathbf{M}_R$$

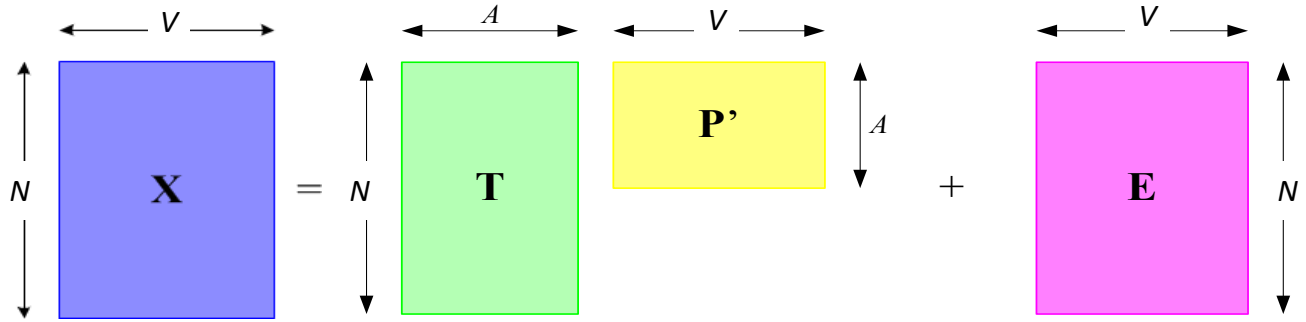


$$R = \text{rank}(\mathbf{X})$$
$$\text{rank}(\mathbf{M}_i) = 1$$

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_R\mathbf{p}_R^{\mathrm{T}}$$



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



Geometrical interpretation of PCA

- **Scores** $\mathbf{T}$: are the projections of the observations in the space of the latent variables (i.e., the coordinates in the PC space)
  - orthogonal and independent
  - **identify the relation among observations**

orthogonality

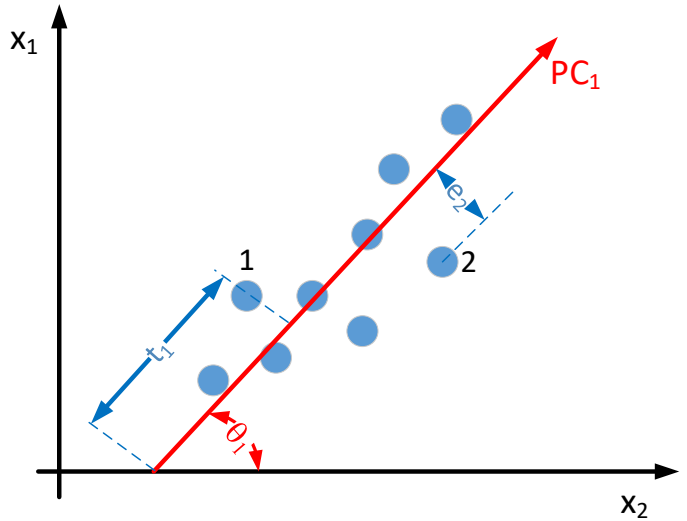
$$\mathbf{t}_n^{\mathrm{T}}\mathbf{t}_n = (N-1)\lambda_n$$
$$\mathbf{t}_n^{\mathrm{T}}\mathbf{t}_j = 0$$

- **Loadings** $\mathbf{P}$: are the director cosines of the PCs
  - identify the direction of maximum variability of the data
  - orthonormal
  - eigenvectors of the covariance/correlation matrix of $\mathbf{X}$
  - **identify the correlation between variables**

ortho-normality

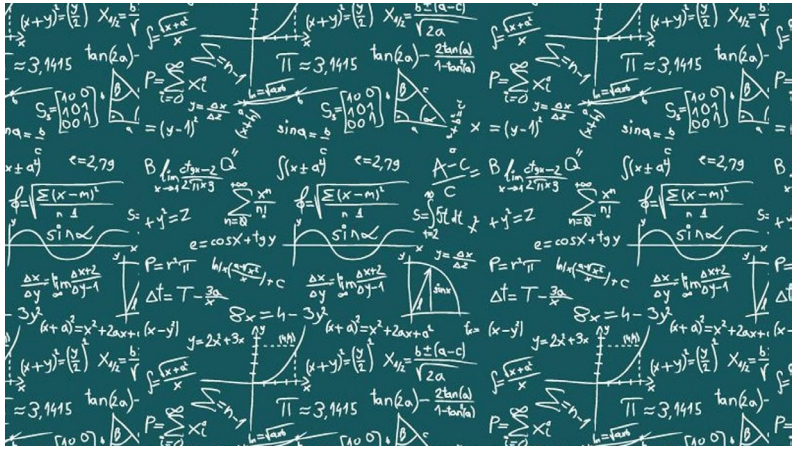
$$\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_n = 1$$
$$\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_j = 0$$

- **Residuals** $\mathbf{E}$: 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 PC space)



PCA tasks

- PCA is an **unsupervised correlative latent-variables methodology** which performs the following **tasks**:
  - from the mathematical point of view:
    - exploratory analysis
    - data mining
    - dimensionality reduction
    - correlative analysis
      - correlation between V variables
      - relation between N observations
    - clustering, etc…
  - from the engineering point of view:
    - process analytical technologies exploitation
    - process understanding
    - process troubleshooting
    - predictive maintenance
    - process monitoring
    - root cause diagnosis of anomalies, etc.



Starting points for self-assessment

- Could you remember at least 4 of the main characteristics of a PCA model?
- What is a principal component?
- What are scores?
- What are loadings?
- What are residuals?
- Could you give a geometrical interpretation of PCA?
- Do you have an idea of what could be some applications of PCA?

Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lessons #6 – Part 2

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab

Data pretreatment

Data pretreatment

- Data can have substantially different measurement scales:
  - example:
    - temperature may vary in a range of -100 – 2000°C
    - concentration vary in a range 0-1
    - pressure may vary in the range 1 – 6000 bar
- Variables with large values in their measurement scale seem to have larger variance than variables with low values in the respective scale
- Multivariate statistical techniques model the largest variance in the data

- We want to avoid that the variability we extract through PCA is affected by the scale of difference variables
  - variables must be **normalized**!

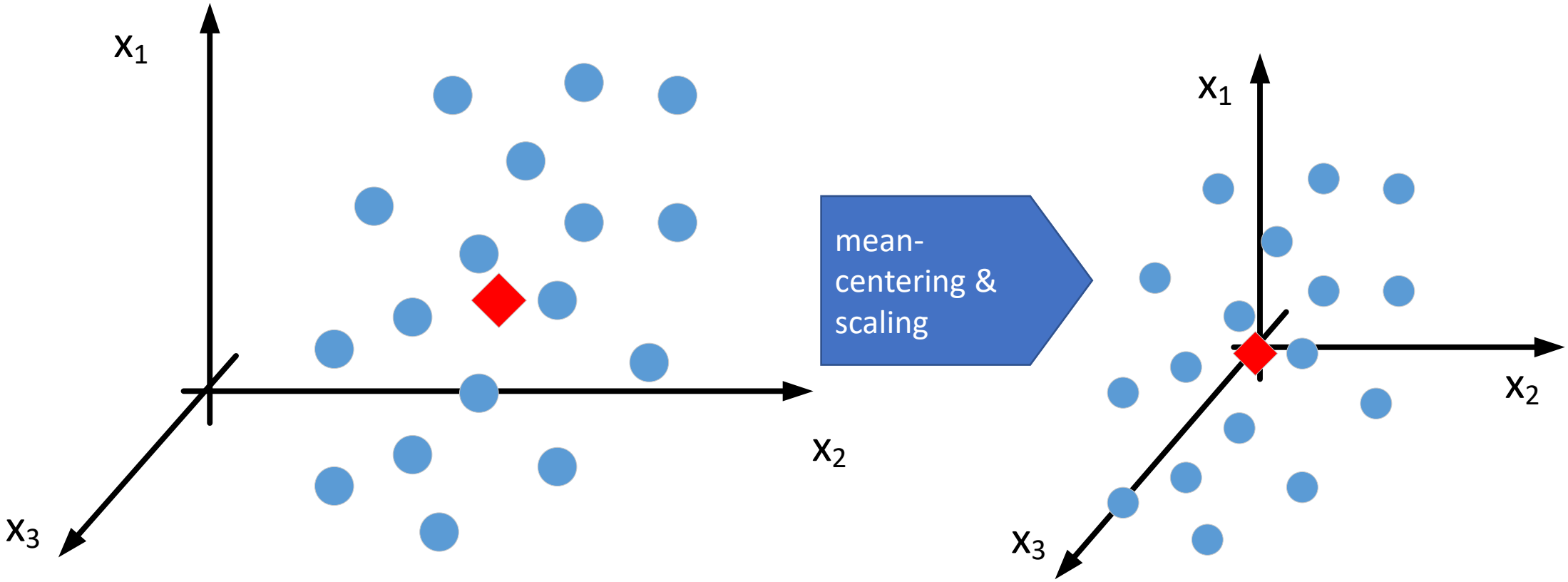
Auto-scaling

- Pre-treatments:
  - depend on:
    - characteristics of the available data
    - objectives of the analysis
  - may include: filtering, denoising, transformations, advanced scaling, data compression
- **Variables must be weighted in a similar way** to exploit the PCA model and to understand their importance
  - **Auto-scaling**

$$\frac{x_{n,v} - \mu_v}{\sigma_v}$$

    - **mean-centering**: subtracting to each element $x_{n,v}$ of $\mathbf{X}$ the mean value of its column μ_v
      - essential for the correct interpretation of the PCA model
      - if not mean-centered, principal components may identify as significant directions of variability in the data the differences between the variable mean values
    - **scaling to unit variance**: dividing each element $x_{n,v}$ of $\mathbf{X}$ by the standard deviation of its column σ_v
      - essential step to make the analysis independent of the units of the variables
      - allows the simultaneous analysis of quantities which have different magnitudes
      - partially linearize data

Geometrical interpretation of autoscaling



The importance of autoscaling data

- Variables can undergo further scaling or weighting operations to determine a different impact of each variable on the model
  - if $\mathbf{X}$ is mean-centered only, matrix $\mathbf{\Sigma} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ represents the **covariance matrix** of $\mathbf{X}$
  - if data are auto-scaled, $\mathbf{\Sigma}$ becomes the **correlation matrix** of $\mathbf{X}$

- **Correlations between variables can be identified from the loadings of a PCA model performed on auto-scaled data!**

Warning on pretreatments

- Pay attention that the most appropriate methodology for data pretreatment depends on the **application** and the considered **type of data**

- For example:
  - to deal with **spectroscopic data**:
    - **standard normal variate** (i.e., autoscaling in the sense of the rows) plus **first and second derivatives** and **baseline correction** are often suggested
  - to deal with **–omics data**:
    - **Pareto scaling** is often suggested
  - etc…

PCA model structure

Selection of the appropriate number A of PCs

Selection of the number of latent variables

- Determination of the dimensionality of the latent space of the model
  - what is the most appropriate number A of PCs to be retained in the PCA model?

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

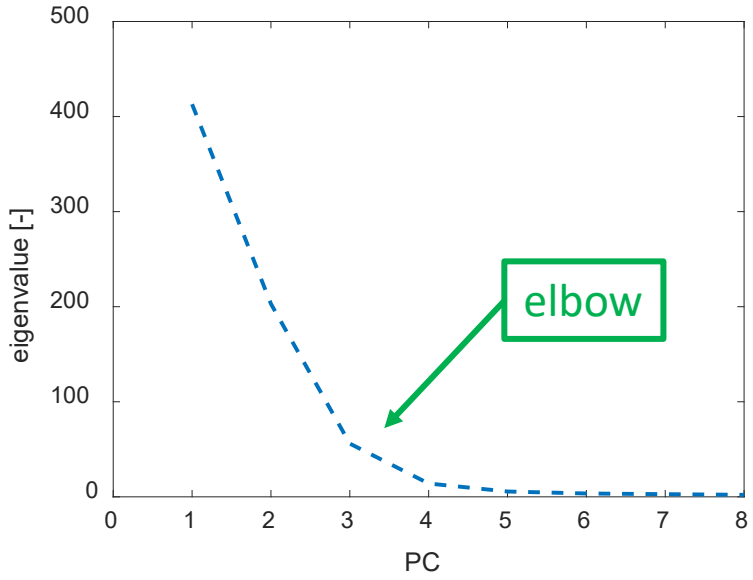
- Different issues should be considered:
  - number of samples
  - total variance explained and relative size of the eigenvalues
    - the variance explained per component
  - subject-matter interpretations of the PCs
- Methods for determining the most appropriate number of PCs
  1. **scree test** (Jackson *et al.*, 1991)
  2. **eigenvalue-greater-than-one rule** (Mardia *et al.*, 1979)
  3. **cross-validation** (Wold, 1978)

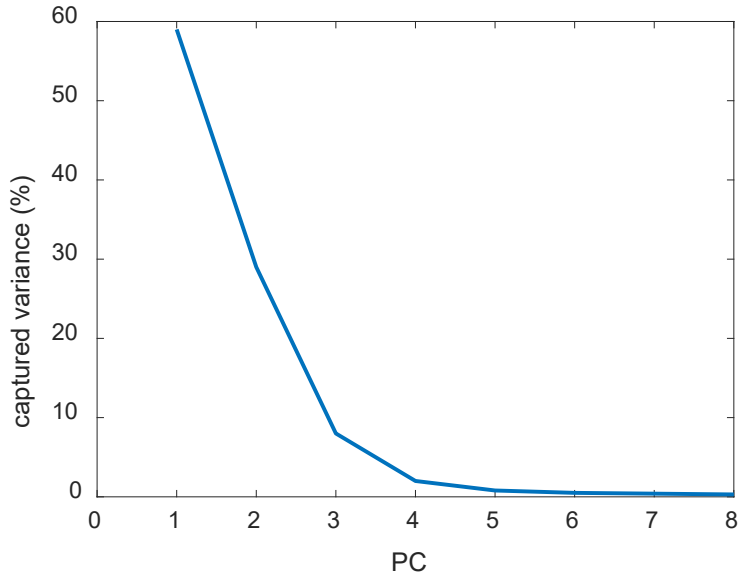
1. Scree test (1/2)

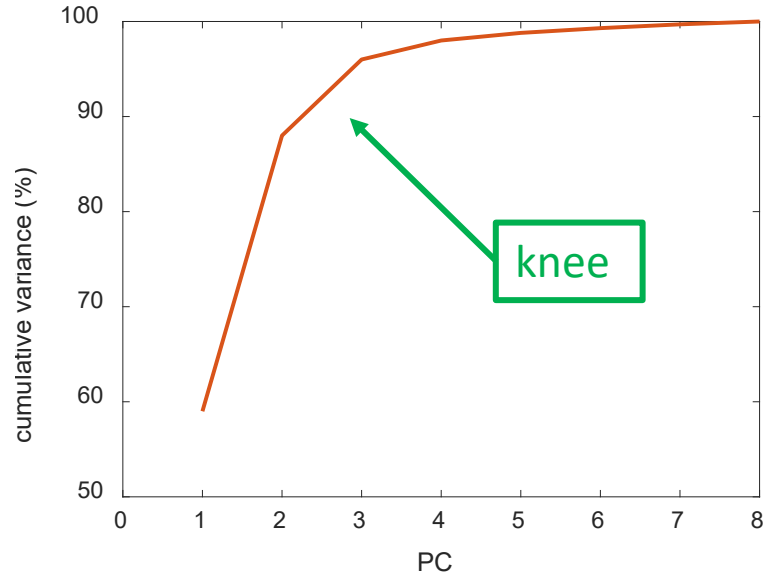
- **Scree test** (Jackson *et al.*, 1991)
  - empirical graphical procedure
  - **analysis of the profile of an index indicating the variability of the original data captured by the PCA model per PC**
    - explained variance R^2 (will be defined in few slides) per PC
    - eigenvalues or the residual percent variance
- Basic idea:
  - the variance described by the model should reach a "steady-state" adding PCs that describe the variability due to random errors
  - the number of PCs to be included in the model is when:
    - a break point is found in the curve (sometimes called "elbow" or "knee")
    - the profile stabilizes
- Advantages:
  - easy
- Drawbacks:
  - if the curve decreases smoothly it can be difficult to identify an "elbow" on it

1. Scree test (2/2)

- Visual inspection of the variance captured by each PC in terms of:
  - eigenvalue
  - captured variance
  - cumulative captured variance

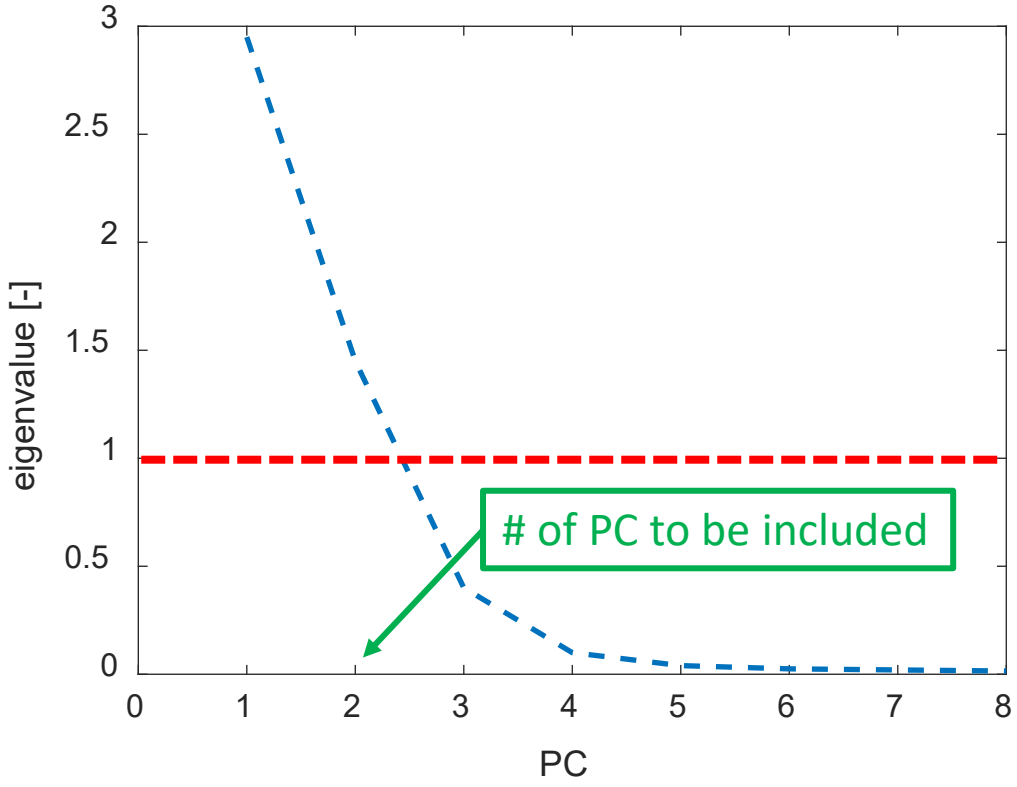






2. Eigenvalue-greater-than-one rule

- **Eigenvalue-greater-than-one rule** (Mardia *et al.*, 1979)
  - all the PCs whose corresponding eigenvalues are lower than one are not considered in the model
- Basic idea:
  - if data are auto-scaled, the eigenvalue corresponding to a PC represents roughly the number of original variables represented by this PC
  - a PC capturing less than one original variable should not be included in the model
- Advantages:
  - very easy to implement and automate
- Drawbacks:
  - in some cases, PCs are discarded even if their eigenvalue is very close to one and their contribution to explain the systematic variability is significant
  - it may be reasonable to lower the threshold in order to include PCs whose eigenvalue may be lower than one ($\sim 0.7 - 0.8$).



3. Cross validation

- **Cross-validation based** on the prediction error sum of squares (Wold, 1978)
- Basic idea:
  - the most appropriate number of PCs to build the model is the one that minimizes the error of reconstruction for new (unknown) samples
- Iterative procedure is repeated by increasing the number of PCs used to build the model (different algorithms can be employed):
  1. divide the $\mathbf{X}$ dataset in G subgroups $\mathbf{X}_g$ of C samples (with $g = 1, \dots, G$)
  2. delete the samples in one of the $\mathbf{X}_g$ groups from the original dataset $\mathbf{X}$
  3. build a PCA model with the reduced dataset
  4. project the data $\mathbf{X}_g$ in the model built in step 3
  5. compute *PRESS* for the reconstruction of $\mathbf{X}_g$
  6. go back to step 1 to select the next subset until all the G subsets have been considered
- The selected number of PCs is the one that minimizes the error, namely the value of *PRESS*
- Advantages:
  - high reliability
- Drawbacks:
  - computationally intensive
  - less reliable when autocorrelation or nonlinearities are present in the data

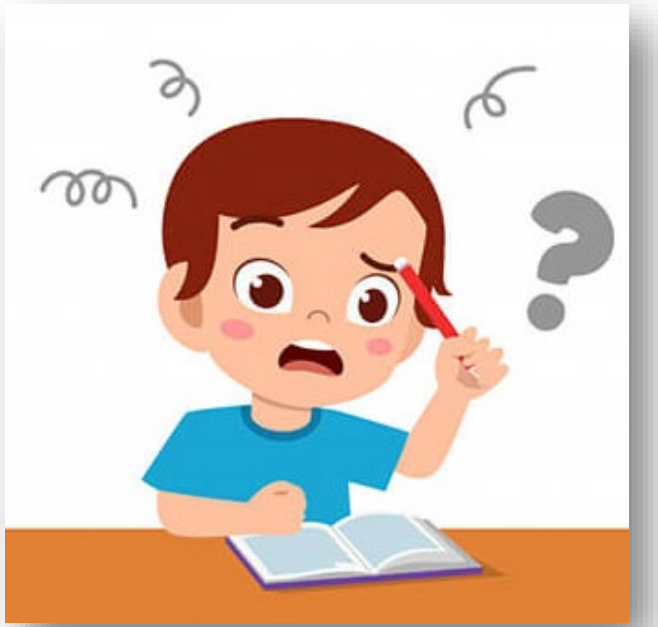
$$PRESS = \sum_{n=1}^{N} \sum_{v=1}^{V} e_{n,v}^2$$

PCA model adequacy

Model and sample diagnostics

PCA diagnostics

- Two basic information are now missing:
  - is a PCA model appropriate to represent the original data with few PCs? Do PCs fit well the original data?
  - how may observation be judged within a PCA model? How do observations conform to the value of the other data and their correlation structure?



PCA model diagnostics

- Several diagnostics can be used to evaluate the performance of a PCA model:
  - model diagnostics
  - variable diagnostics
  - sample diagnostics (Eriksson *et al.*, 2001)
- **Model diagnostics**
  - **coefficient of determination**: the amount of variability of the original data explained by the model (in calibration)

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

    - where $\hat{x}_{n,v}$ is the $[n, v]$ element of $\hat{\mathbf{X}}$ and $\bar{x}_v$ is the mean of variable v
  - the **Q^2 index**: a measure of the **predictive power** of the model (validation)

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

  - 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 of PCs

PCA sample diagnostics

- Sample diagnostics:
  - the **Hotelling's T^2 statistic** (Hotelling, 1933) measures the overall distance of the projections of an observation from the PC space origin
    - since each PC explains a different data variance aliquot, the Mahalanobis distance is used (Mardia *et al.*, 1979):

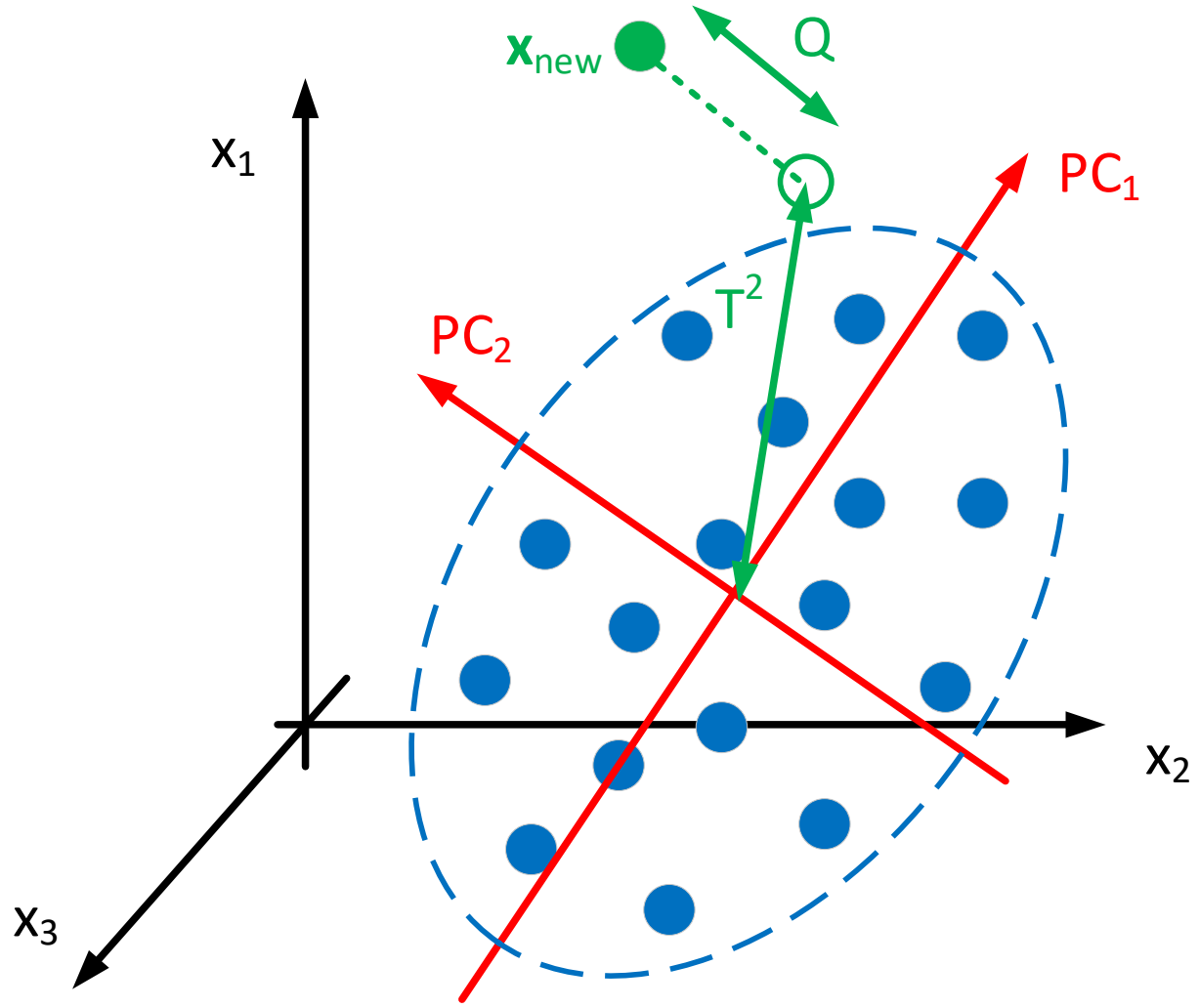
$$T_n^2 = \mathbf{t}_n^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \mathbf{t}_n = \sum_{a=1}^{A} \frac{t_{a,n}^2}{\lambda_a}$$

      - where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues
      - the T^2 statistic represents the multivariate generalization of the Student's t-test, and provides a check for observations adhering to **multivariate normality**
  - 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

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

      - it accounts for the mismatch of the model in representing the n-th observation

Geometrical interpretation of the PCA diagnostics



PCA in Matlab®

- PCA can be managed with several commands in Matlab®:
  - `pca` (PLS_Toolbox, Eigenvector Research Inc.)
  - `princomp`
  - `svd`
    - this requires calculating:
      - the loadings $\mathbf{P}$ as eigenvectors of the covariance matrix of $\mathbf{X}$
      - the scores from $\mathbf{P}$ and $\mathbf{X}$
  - Very easy code…

- For the sake of simplicity, we will use mainly the PLS_Toolbox® graphic user interface

Starting points for self-assessment

- Could you remember at least 2 effective methods to select the correct number of PCs to be selected to build a PCA model?
- How could you normalize data? What is data normalization intended for?
- Could you give a definition of what is Q^2 and how it is related to R^2?
- What is T_n^2 used for? And Q_n?

… per sempre a fianco a me!