2.1 Data analytics for dimensionality reduction: Principal Component Analysis (PCA)

Prof. Massimiliano Grosso
University of Cagliari, Italy
massimiliano.grosso@dimcm.unica.it
GRICU PhD School 2021
Digitalization Tools for the Chemical and Process Industries
March 12, 2021

1

Outline

- Motivations
- Basic concepts
- Preprocessing
- Mathematical background
- Dimension reduction
- Geometrical interpretation

2

Motivations

- Concerns when dealing with **“huge” amount of data**:
  - The **size** of the data:
    - The **useful** information is often «**hidden**» amongst **hundred/thousands of variables**
  - The measurements are often **highly correlated** with one another (multicollinearity)
    - The number of **independent variables** (**degrees of freedom**) is **much less** than the number of measurements on hand
- **Noise** in the measurements
  - Difficulties in distinguishing the noise from the deterministic variations induced by external sources

3

Motivations

- Multivariate data analysis method for
  - Explorative data analysis
  - Outlier detection
  - Rank reduction
  - Graphical clustering
  - Classification
- PCA allows interpretation based on **all** variables simultaneously, leading to understanding deeper than what is possible looking at the individual variables alone
- It is the first **multivariate analysis** to be carried out

4

PCA: Basic concepts

- Aim of the PCA:

Original variables → **Projection of the variables onto the Principal components (PCs)**

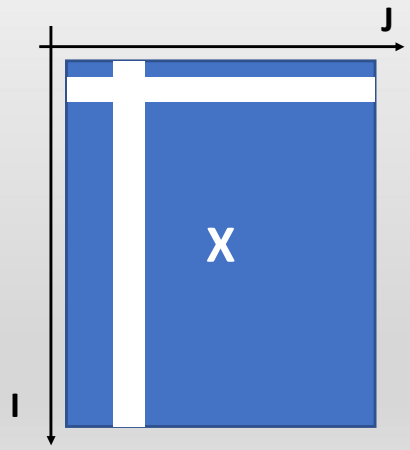
high dimension, strongly correlated

artificial variables, dimension **much lower**, **independent**

5

PCA: Basic concepts

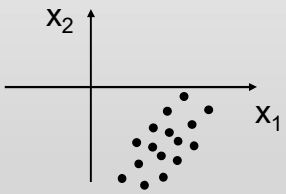
- Data must be collected on matrix **X**
- Column vectors represent the **variables (j=1,…,J)**
  - attributes, wavelenghts, physical/chemical parameters etc.
- Row vectors represent the **samples (i=1,…,I)** collected during the experiments



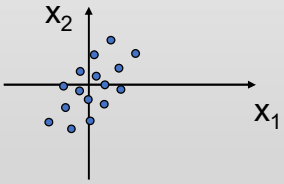
6

Preprocessing of the data

- Matrix **X** can be visualized in a coordinate system made up by
  - J orthogonal axes, each representing one of the original J variables
  - Each i-th sample is a J-dimensional row vector
- Two-dimensional example with two variables highly correlated



- First step;
- **Translate** the data to the center («**mean centering**»)



$$x_{ij} = x^*_{ij} - \bar{x}^*_{\bullet j} \quad where\ \bar{x}^*_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} x^*_{ij}$$

7

Preprocessing of the data

- Mean centering allows to consider the covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{bmatrix} = \begin{bmatrix} x^*_{11} - \bar{x}^*_{\bullet 1} & \cdots & x^*_{1j} - \bar{x}^*_{\bullet j} \\ \vdots & \ddots & \vdots \\ x^*_{I1} - \bar{x}^*_{\bullet 1} & \cdots & x^*_{IJ} - \bar{x}^*_{\bullet j} \end{bmatrix}$$

- Indeed, for the element *kl*

$$\mathbf{C_{kl}} = (\mathbf{X}^T\mathbf{X})_{kl} = \sum_{i=1}^{I} (x^*_{ik} - \bar{x}^*_{\bullet k})(x^*_{il} - \bar{x}^*_{\bullet l}) = \sum_{i=1}^{I} x_{ik}x_{il}$$

- The diagonal elements of **C** are the *dispersion* related to the j-th variable

$$\mathbf{C_{jj}} = (\mathbf{X}^T\mathbf{X})_{jj} = \sum_{i=1}^{I} \left(x^*_{ij} - \bar{x}^*_{\bullet j}\right)^2 = \sum_{i=1}^{I} x^2_{ij}$$

8

PCA – Basic concepts

- Principal Component Analysis is based on the **decomposition** of the dataset matrix **X**

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}}$$

(I×J) (I×J)(J×J)

Scores matrix
Artificial variables generated by PCA

Loadings matrix
Rotation matrix relating artificial variables with the original ones

9

PCA – Basic concepts

- Important properties:

1. Even the scores are mean centered

$$\bar{x}_{\bullet j} = 0 \;\; \forall j = 1, \dots, J \quad \Rightarrow \bar{t}_{\bullet j} = 0 \qquad \forall j = 1, \dots, J$$

2. **Column vectors of the score matrix T are orthogonal**: $\mathbf{t}_m^T \mathbf{t}_n = 0 \; \forall m \neq n$
   1. The square of the score matrix $\mathbf{\Lambda} = \mathbf{T}^{\mathbf{T}} \cdot \mathbf{T}$ is **diagonal**
3. **Loadings matrix P is orthogonal:** $\mathbf{P}^T = \mathbf{P}^{-1} \Rightarrow \mathbf{P}^T \cdot \mathbf{P} = \mathbf{I}$

10

Mathematical background

- PCA scores and loadings can be related to the computation of the **eigenvalues** and **eigenvectors** of the J×J **covariance** matrix

$$\mathbf{C} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$$

- **Remark**
- **C** is a square, symmetric matrix, this leads to the following properties:
  - All the eigenvalues are real and positive
  - All the **eigenvectors** are **orthogonal** to each other

11

Mathematical background

- Starting from the definition $\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\boldsymbol{T}}$, one can obtain the following relationships

$$\mathbf{C} = \mathbf{X}^{\mathbf{T}} \cdot \mathbf{X} = \mathbf{P} \cdot \mathbf{T}^{\mathbf{T}} \cdot \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} = \mathbf{P} \cdot \boldsymbol{\Lambda} \cdot \mathbf{P}^{T} = \mathbf{P} \cdot \boldsymbol{\Lambda} \cdot \mathbf{P}^{-1}$$

- The latter equation corresponds to the **eigendecomposition** of the square matrix $\mathbf{C} = \mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}$
  - $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of **C**
  - The m-th element $\mathbf{t}_m^T \mathbf{t}_m = \boldsymbol{\Lambda}_{mm}$ is the **variance** explained by the m-th score
  - **P** is the *n*×*n* square matrix whose *m*-th column is the eigenvector $\mathbf{p}_m$ of **C**
  - **it is a rotation matrix**

12

Mathematical background

- Once the eigenvectors $\mathbf{p}_m$ are computed the corresponding scores can be derived

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^T \Rightarrow \quad \mathbf{X} \cdot \mathbf{P} = \mathbf{T} \cdot \mathbf{P}^T \cdot \mathbf{P} \Rightarrow \quad \mathbf{T} = \mathbf{X} \cdot \mathbf{P}$$

- In practice, the original variables are **projected** onto the orthogonal eigenspace defined by the eigenvectors/loadings

Mathematical background

- The eigenvalues of the covariance matrix are related to the variance of the scores

$$\lambda_j = \mathbf{t}_j^T \mathbf{t}_j = \sum_{i=1}^{I} t_{i,j}^2$$

- Thus the j-th eigenvalue is the **dispersion captured by the j-th score**
- **The total variance in the original data set is preserved in the T matrix**

Sum of the variances of the original variables = **Sum of the variances of the scores**

Mathematical background

- In summary, one ends up with two matrices

$$\underset{(I \times J)}{\mathbf{T}} = \left[\underset{(I \times 1)}{\mathbf{t}_1}, \underset{(I \times 1)}{\mathbf{t}_2}, \cdots \underset{(I \times 1)}{\mathbf{t}_J}\right] \qquad \underset{(J \times J)}{\mathbf{P}} = \left[\underset{(J \times 1)}{\mathbf{p}_1}, \underset{(J \times 1)}{\mathbf{p}_2}, \cdots \underset{(J \times 1)}{\mathbf{p}_J}\right]$$

Scores matrix

The j-th column represents an **independent variable** obtained by projecting the data onto the j-th eigenvector

Remind:

Sort the eigenvectors according to their eigenvalue size (that is, their variance)

Loading

Each column is an eigenvector of the covariance matrix

PCA – Dimension reduction

- The scores and loading matrices can be approximated by considering only the **first *A* principal components**

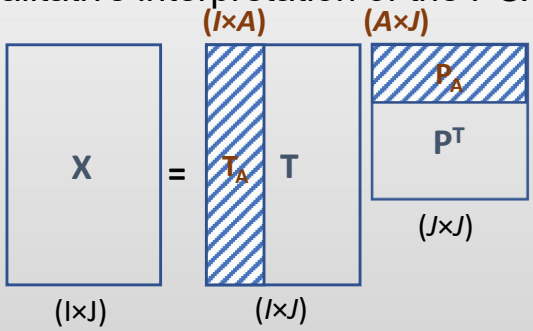
$$\underset{(I \times I)}{\mathbf{T}} = \left[\underset{(I \times A)}{\mathbf{T}_A} \vdots \underset{(I \times (I-A))}{\tilde{\mathbf{T}}_A}\right] \approx \underset{(I \times A)}{\mathbf{T}_A} \qquad \underset{(J \times I)}{\mathbf{P}} = \left[\underset{(J \times A)}{\mathbf{P}_A} \vdots \underset{(J \times (I-A))}{\tilde{\mathbf{P}}_A}\right] \approx \underset{(J \times A)}{\mathbf{P}_A}$$

Information considered negligible

Information considered negligible

PCA – Dimension reduction

- Qualitative interpretation of the PCA



$$\underset{(I\times J)}{\mathbf{X}} \approx \underset{(I\times A)}{\mathbf{T}_A}\underset{(A\times J)}{\mathbf{P}_A^T}$$

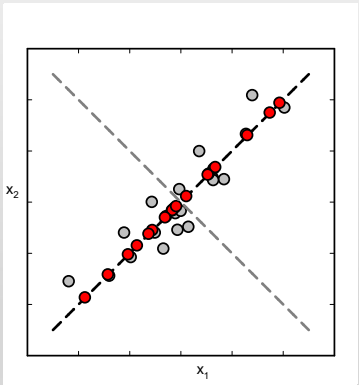
In general:
$A \ll J$

- Only part of the information collected in the **X** matrix is relevant
- Only the first *A* columns of **T** (the first scores) take into account most of the data variance

17

PCA – A geometrical interpretation

- 2D example - Reduction to 1D

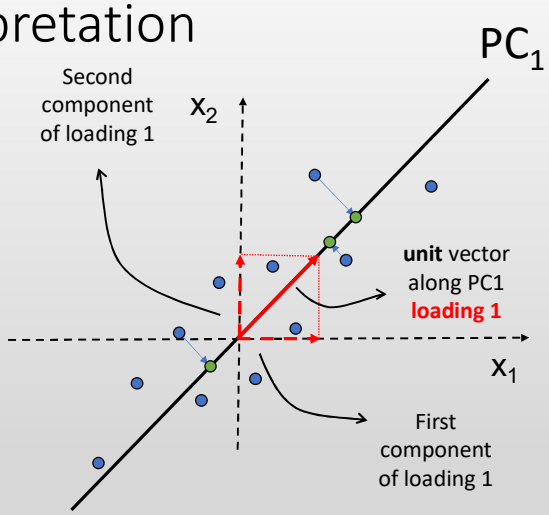


- Samples are strongly correlated
- First principal component PC1 is the eigenvector direction corresponding to maximum variance (largest eigenvalue) in the coordinate space
- Second principal component is the orthogonal one leading to the second variance directions

18

PCA – A geometrical interpretation

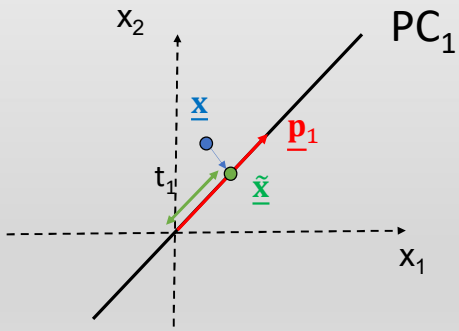
- **Orthogonal projection** onto a specific PC results in a **score** for **each sample**
- The loading is the **unit vector** which defines this direction



19

PCA – A geometrical interpretation

- The score is the projection of the point onto the **first** principal component



$$\underline{\mathbf{x}} \approx \underline{\tilde{\mathbf{x}}} = t_1 \mathbf{p_1}$$

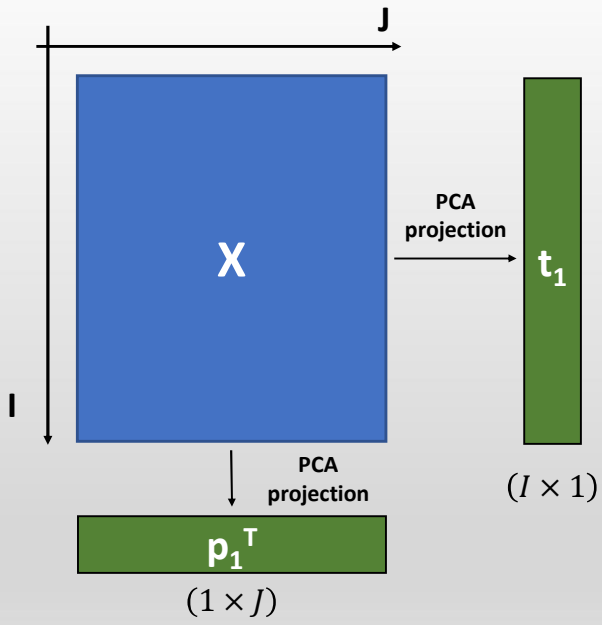
20

PCA – Working principle – Reduction to 1D

- PCA projects matrix **X** into:
  - a **score** vector $\mathbf{t}_1$
  - a **loading** vector $\mathbf{p}_1$

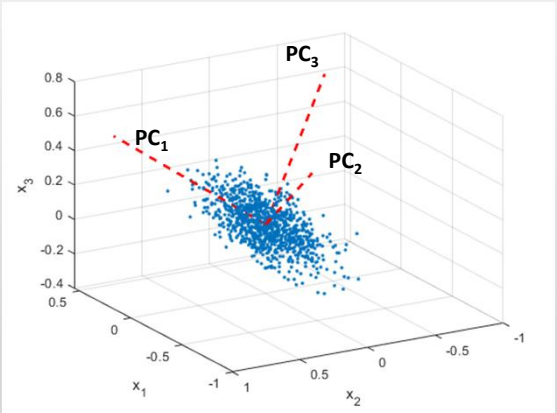
$$\underset{(I \times J)}{\underline{\underline{X}}} \approx \underset{(I \times 1)}{\underline{\mathbf{t}_1}} \; \underset{(1 \times J)}{\underline{\mathbf{p}_1^T}}$$

- $\mathbf{t}_1$ and $\mathbf{p}_1$ are the first components



21

PCA – A geometrical interpretation

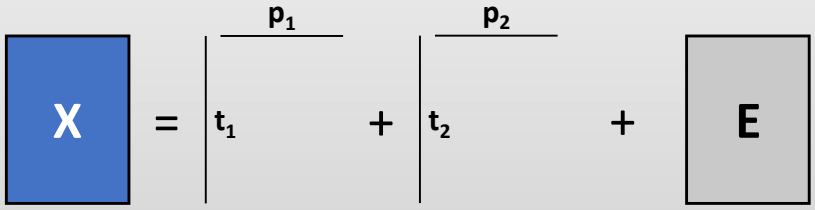


- 3D example (A little bit more complicated)
- Points are mostly aligned along the 2D plane defined by the PC_1 and the PC_2 directions

22

PCA – Working principle – Reduction to 2D

- If two principal components are required, matrix is formed by the outer products of $\mathbf{t}_1$ and $\mathbf{p}_1$, $\mathbf{t}_2$ and $\mathbf{p}_2$

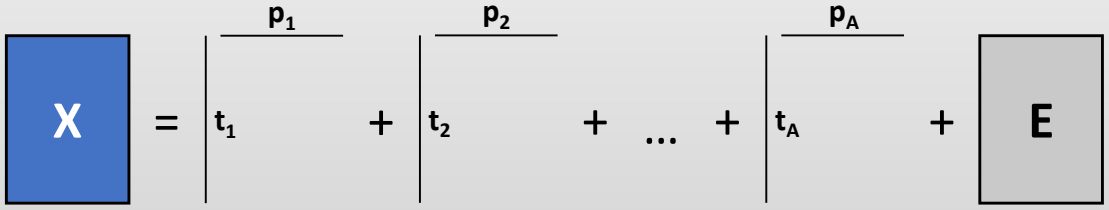


- Matrix **X** is decomposed into two sets of rank 1 outer products (2 terms) and the **residual matrix E**

23

PCA – Working principle

- Successive components are formed by the outer products of $\mathbf{t}_a$ and $\mathbf{p}_a$



- Matrix **X** is decomposed into a set of rank 1 outer products (A terms) and the **residual matrix E**

24

PCA – Working principle

- The master equation for PCA is eventually

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^t + \mathbf{t}_2\mathbf{p}_2^t + \ldots \mathbf{t}_A\mathbf{p}_A^t + \mathbf{E}$$

- or

$$\underset{(I \times J)}{\boldsymbol{X}} = \underset{(I \times A)}{\mathbf{T}_A}\ \underset{(A \times J)}{\mathbf{P}_A^T} + \underset{(I \times J)}{\mathbf{E}}$$

original data matrix — **score** matrix — **loading** matrix — **residual** matrix

25

Estimation of the residuals

- When considering a PCA model with *A* principal components, one can evaluate the residual **E**

$$\mathbf{E} = \mathbf{X} - \mathbf{T}_A \cdot \mathbf{P}_A^T = \mathbf{X} \cdot \left(\mathbf{I} - \mathbf{P}_A \cdot \mathbf{P}_A^T\right)$$

26

Estimation of the components

- How many principal components are needed?
- Possible criterion: **cumulative variance** explained by the **first *A*** principal components
  - The number of principal components to be considered explains most of the variance in the data (e.g., 95%)
- Alternative possibilities will be discusses in the case studies

27

PCA to predict new data – Projection of the data onto the principal component space

- Single observations, (eventually new data $\mathbf{x}_{new}$) can be eventually **projected** onto the space defined by the PCA model:

$$\underset{(1\times A)}{\mathbf{t}_{new}} = \underset{(1\times J)}{\mathbf{x}_{new}} \cdot \underset{(J\times A)}{\mathbf{P}_A}$$

$$\underset{(1\times J)}{\hat{\mathbf{x}}_{new}} = \underset{(1\times J)}{\mathbf{x}_{new}} \cdot \underset{(J\times A)}{\mathbf{P}_A} \cdot \underset{(A\times J)}{\mathbf{P}_A^T}$$

28

PCA – Summary

- PCA projects the original data onto an orthogonal eigenspace of smaller dimensions
- The space is described by the first *A* eigenvectors of the covariance matrix
- The scores (i.e. the data projections onto the first eigenvectors) represent a set of independent variables
- New data can be projected in the PCA model

29

References

1. Brereton, R.G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* Wiley, 2003
2. Brereton, R.G. *Chemometrics for Pattern Recognition.* Wiley, 2009
3. Jackson, J.E., *A User's Guide to Principal Components*. Wiley, New York, 1991
4. Jolliffe, I.T. *Principal Component Analysis*. Second Edition. Springer, 2002.
5. Jolliffe IT, Cadima J. (2016). Principal component analysis: a review and recent developments. *Phil.Trans.R.Soc.* A374:20150202.
6. Wold S., Esbensen K.,Geladi P (1987). Principal Component Analysis – A tutorial. *Chemom. Intell. Lab.* 2, 37-52

30