From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

Principal Component Analysis

- Introduction

➢ **Definition**

- the principal components of a collection of points in a real p-space that are a sequence of p direction vectors, where the i^{th} vector is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vectors
- a best-fitting line is defined as one that minimizes the average squared distance from the points to the line
- these directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated
- → PCA: process of computing the principal components and using them to perform a change of basis on the data

Principal Component Analysis

- Introduction

➢ **Uses**

○ **principal component analysis** (PCA), also known as the *Karhunen-Love* transform, is widely used for:

❑ **dimensionality reduction**:
it projects data points living in a d-dimensional space onto a M-dimensional subspace, where $M < d$

○ if $M = 2$, PCA allows **data visualization** while preserving the variance of the original data

❑ **feature extraction**: it generates new uncorrelated (i.e., without redundancies) meaningful features

Principal Component Analysis

- Introduction

➢ Example

- main characteristics of the planets of the solar system:
  - distance to the sun (in UA)
  - diameter (in km)
  - density (in g / cm^3)

	Distance	Diamètre	Densité
Mercure	0,387	4 878	5,42
Vénus	0,723	12 104	5,25
Terre	1,000	12 756	5,52
Mars	1,524	6 787	3,94
Jupiter	5,203	142 800	1,31
Saturne	9,539	120 660	0,69
Uranus	19,180	51 118	1,29
Neptune	30,060	49 528	1,64
Pluton	39,530	2 300	2,03

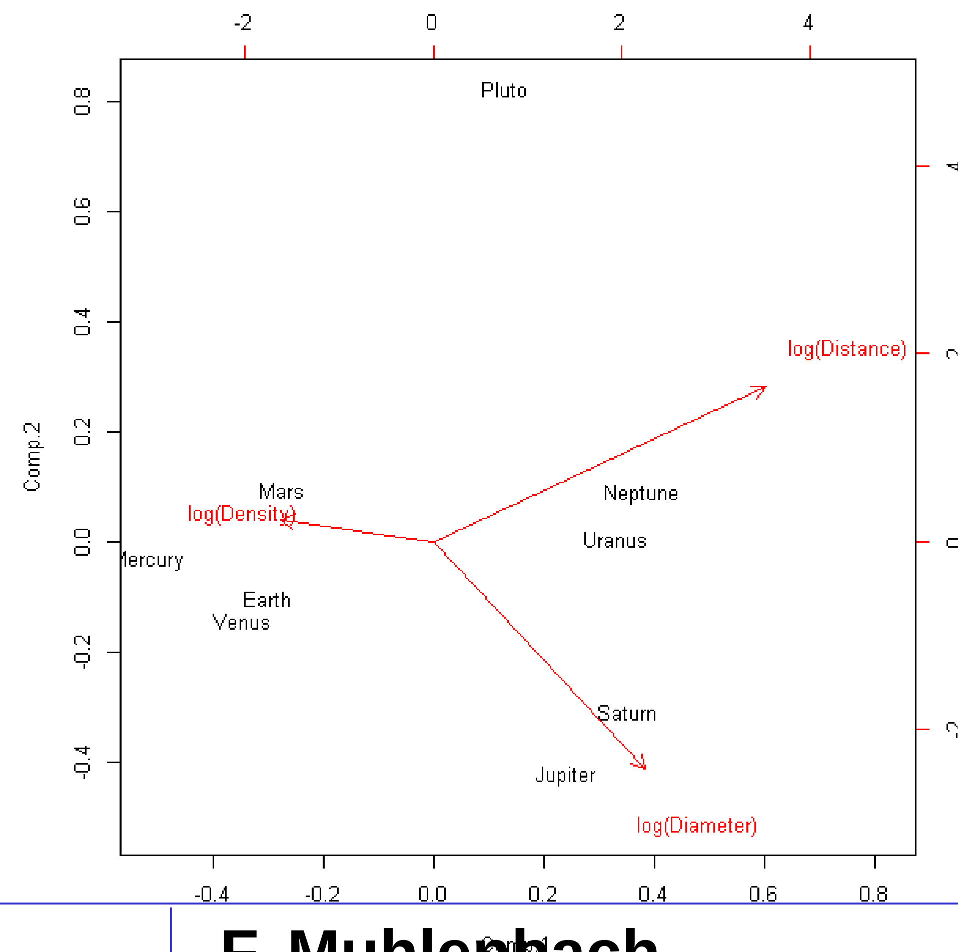
- since a 3D-plot is not always very readable, can we find a 2D-plot of the data such that close points in that new space mean similar planets in the original 3D-space?

Principal Component Analysis

- Introduction

➢ **Example**

○ possible solution with PCA: reduction of dimensionality

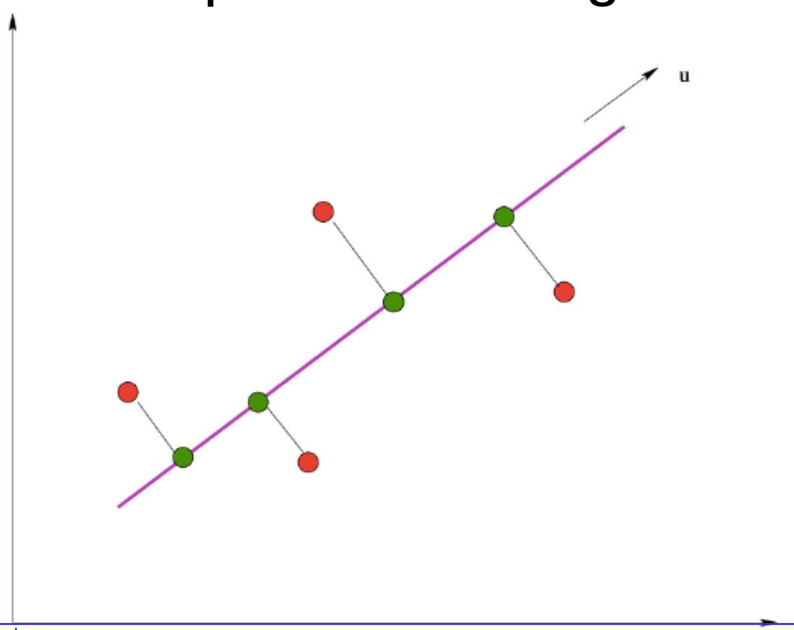


• Goal of PCA

Example

- the goal of PCA is to linearly project the data $\boldsymbol{x}_i \in \mathbb{R}^d$ onto a space having dimensionality $M < d$ such that close points in that new M-space mean similar examples in the original d-space

- here, $d = 2$ and $M = 1$
- we have to define the direction of this space using a 2-dimensional vector **u**



- Maximization of the variance of the projected data

- ○ let us suppose that the training data are zero mean (that is, $\forall i$, x_i is changed into $x_i \leftarrow xi - \bar{x}$)
- ○ PCA seeks a new space of size $M < d$ by applying a linear transformation $\mathbf{U}^{\mathrm{T}}$ on the original data
- ○ the new representation of a training data x_i, denoted by t_i, is computed as follows: $t_i = \mathbf{U}^{\mathrm{T}} x_i$
- ○ where $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_M)$ is a $(d \times M)$-matrix of new bases and $\mathbf{u}_j \in \mathbb{R}^d$
- ○ we impose that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = I$, that is $\mathbf{U}$ is orthogonal, meaning:
  - ❑ every new feature $\mathbf{u}_i$ is linearly independent from the others,
  - ❑ $\forall j$, $\mathbf{u}_j{}^{\mathrm{T}}\mathbf{u}_j = 1$
- ○ note that each t_i is a linear combination of the original features

Principal Component Analysis

- Maximization of the variance of the projected data
  - if $x_j \in \mathbb{R}^d$, then the PCA can generate a maximum of d new components, i.e., $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_d)$ is a $(d \times d)$-matrix of new bases
  - if the linear transformation $\mathbf{U}$ is composed with d new bases, then it is possible to perfectly rebuild the data of the initial space (i.e., it is a bijection) but if $\mathbf{U}$ is composed only with M new bases, with $M < d$, then some information is lost in the projection in the M-dimension space and the reconstruction of the data in the initial space is not perfect
  - therefore the objective of the PCA is to minimize this reconstruction error in order to keep as much information as possible from the original space despite the dim. reduction

Principal Component Analysis

- Maximization of the variance of the projected data

- let $\hat{x}_i = \mathbf{U}t_i$ be the reconstruction of the original vector x_i using the transformation $\mathbf{U}$
- the objective of PCA is to optimize $\mathbf{U}$ s.t. the mean square error $J(\mathbf{U})$ between x_i and $\hat{x}_i$ is as small as possible:

$$\min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i (x_i - \hat{x}_i)^2$$

$$\Leftrightarrow \min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i (x_i - \mathbf{U}\mathbf{U}^\top x_i)(x_i - \mathbf{U}\mathbf{U}^\top x_i)$$

$$\Leftrightarrow \min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i (x_i{}^\top x_i - 2x_i{}^\top \mathbf{U}\mathbf{U}^\top x_i + x_i{}^\top \mathbf{U}\mathbf{U}^\top \mathbf{U}\mathbf{U}^\top x_i)$$

$$\Leftrightarrow \min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i (x_i{}^\top x_i - x_i{}^\top \mathbf{U}\mathbf{U}^\top x_i) \text{ because } \mathbf{U}^\top\mathbf{U} = 1$$

$$\Leftrightarrow \min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i x_i{}^\top x_i - \frac{1}{n} \sum_i x_i{}^\top \mathbf{U}\mathbf{U}^\top x_i$$

• Maximization of the variance of the projected data

➢ Optimization of $\mathbf{U}$ → minimization of $J(\mathbf{U})$ (conclusion):

○ $\min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \frac{1}{n} \sum_i (x_i - \hat{x}_i)^2 = \min_{\mathbf{U}} \frac{1}{n} \sum_i x_i{}^{\mathsf{T}} x_i - \frac{1}{n} \sum_i x_i{}^{\mathsf{T}} \mathbf{U} \mathbf{U}^{\mathsf{T}} x_i$

$\Leftrightarrow \min_{\mathbf{U}} J(\mathbf{U}) = \min_{\mathbf{U}} \mathrm{Tr}(\Sigma) - \mathrm{Tr}(\mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U})$

○ where Σ is the covariance matrix of the original data and $\mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U}$ is covariance in the new space

○ since $\mathrm{Tr}(\Sigma)$ does not depend on $\mathbf{U}$, minimizing $J(\mathbf{U})$ boils down to maximizing $\mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U}$, that is,

$$\max_{\mathbf{U}} \mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U}$$
$$\text{s.t.}\ \forall j = 1, \cdots, M, \mathbf{u}_j{}^{\mathrm{T}} \mathbf{u}_j = 1$$

Principal Component Analysis

- Maximization of the variance of the projected data

➢ Minimization of $J(\mathbf{U})$ → optimization problem:

$$\max_{\mathbf{U}} \mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U}$$
$$\text{s.t. } \forall j = 1, \cdots, M, \mathbf{u}_j^{\mathrm{T}} \mathbf{u}_j = 1$$

- introducing Lagrange multipliers (denoted by the feature vector λ=(λ_1,⋯,λ_M)), we get the unconstrained maximization problem:

$$\max_{\mathbf{U}} \mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U} + \lambda(1 - \mathbf{U}^{\mathrm{T}} \mathbf{U})$$

- let us consider the first component $\mathbf{u}_1$ of the new space
- find $\mathbf{u}_1$ requires to solve:

$$\frac{\partial \mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U} + \lambda(1 - \mathbf{U}^{\mathrm{T}} \mathbf{U})}{\partial \mathbf{u}_1} = 0$$

Principal Component Analysis

- Maximization of the variance of the projected data

➢ **Derivatives of matrices and vectors**

- let $v \in \mathbb{R}^d$ a vector and M a $d \times d$ matrix:

$$\frac{\partial v^{\mathrm{T}} M v}{\partial v} = (M + M^{\mathrm{T}})v$$

- if M is symmetric, then M=M^{T} and

$$\frac{\partial v^{\mathrm{T}} M v}{\partial v} = 2Mv$$

➢ applying the previous on

$$\frac{\partial \mathbf{U}^{\mathrm{T}} \Sigma \mathbf{U} + \lambda(1 - \mathbf{U}^{\mathrm{T}} \mathbf{U})}{\partial \mathbf{u}_1} = 0$$

we get $\Sigma \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$

- which says that $\mathbf{u}_1$ must be an eigenvector of Σ

Principal Component Analysis

- Maximization of the variance of the projected data

- for maximizing $\mathbf{U}^{\mathrm{T}}\Sigma\mathbf{U}$, we have the constraint $\Sigma\mathbf{u}_1 = \lambda_1\mathbf{u}_1$
- if we left-multiply by $\mathbf{u}_1{}^{\mathrm{T}}$ and make use of $\mathbf{u}_1{}^{\mathrm{T}}\mathbf{u}_1 = 1$, we see that the variance is given by

$$\mathbf{u}_1{}^{\mathrm{T}}\Sigma\mathbf{u}_1 = \lambda_1$$

- and the variance will be maximum when we set $\mathbf{u}_1$ equal to the eigenvector having the largest eigenvalue λ_1
- this eigenvector is known as the **first principal component**

- **conclusion**: constraining $\mathbf{U}^{\mathrm{T}}\mathbf{U} = I$ means that we restrict the optimization problem to find an orthogonal matrix $\mathbf{U}$
- therefore, we get the same result $\mathbf{U}^{\mathrm{T}}\Sigma\mathbf{U} = \lambda$ as that of which would have been obtained with a diagonalizable PD matrix

Principal Component Analysis

• Properties of the components

- the eigenvalues of Σ are always positive because is Σ PSD
- the number of components is equal to the number of non zero eigenvalues
- the total variance of the original data is $V = \mathrm{Tr}(\Sigma)$ because the diagonal elements of Σ contain the variances
- we deduce that:

$$V = \mathrm{Tr}(\Sigma) = \mathrm{Tr}(\mathbf{U}\lambda\mathbf{U}^{-1}) = \mathrm{Tr}(\mathbf{U}^{-1}\mathbf{U}\lambda) = \mathrm{Tr}(\lambda) = \lambda_1 + \lambda_2 \cdots + \lambda_d$$

- when we project the data on a two-dimensional plane corresponding to the eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ associated with the two largest eigenvalues λ_1, λ_2, we get a new covariance matrix $\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}$ whose total variance $\hat{V} = \mathrm{Tr}(\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}) = \lambda_1 + \lambda_2$

Principal Component Analysis

- Properties of the components

- projection of data onto a 2-dimensional plane space → covariance matrix $\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}$ with variance $\hat{V} = \mathrm{Tr}(\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}) = \lambda_1 + \lambda_2$
- therefore, we can compute the ratio of variance "explained" by the projected data: $\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\cdots+\lambda_{d-1}+\lambda_d}$
- the higher the ratio, the better the projection
- interpretation of the results of the PCA → it depends on:
- ❑ quality of the representation on the main planes
- ❑ choice of size (number of axes to be used)
- ❑ "internal" interpretation (correlations between variables, place and importance of individuals, size effect, etc.)
- ❑ "external" interpretation (variables and additional individuals)

Principal Component Analysis

- Algorithmic complexity of PCA

- PCA involves evaluating the mean $\bar{x}$ and the covariance matrix Σ of the data set and then finding the M eigenvectors of Σ corresponding to the M largest eigenvalues:
  - the computational cost of computing the full eigenvector decomposition for a matrix of size $d \times d$ is $\mathcal{O}(d^3)$
  - however, if we are only interested in the the projection onto the first M principal components, efficient techniques exist, such as the *power method* that scale like $\mathcal{O}(Md^2)$, or alternatively we can make use of the EM algorithm