Data Science
Static data analysis

Stéphane Marchand-Maillet

Master en Sciences Informatiques - Semestre d'Automne

Representation spaces

Quantitative space

Corresponds to a matrix providing the association

$$X = \text{VARIABLES}\quad \begin{array}{c|cccc} & \multicolumn{4}{c}{\text{ELEMENTS}} \\ \hline & p_1 & p_2 & \cdots & p_M \\ \hline x^1 & \cdots & \cdots & \cdots & \cdots \\ x^2 & \vdots & \ddots & & \\ \vdots & \vdots & & \ddots & \\ x^N & \vdots & & & \ddots \\ \hline \end{array}$$

- ⋆ $N \times M$ matrix of variables x_i^k
- ⋆ Each column contains vector $\mathbf{x}_i$ describing p_i via N variables

on décrit un individu avec N variables

$$\rightsquigarrow \begin{array}{cc} & p_1 \\ x_1 & 1 \\ x_2 & 3 \end{array} \longrightarrow \text{décrivent } p_1$$

Representation space

Categorical values

How to represent **symbolic** values?

$$X = \text{VARIABLES}\ \begin{array}{c|cccc} & \multicolumn{4}{c}{\text{ELEMENTS}} \\ & p_1 & p_2 & \cdots & p_M \\ \hline x^1 & 0 & 1 & \ldots & 1 \\ x^2 & 1 & \ddots & & \\ \vdots & \vdots & & 0 & \\ x^N & 1 & & & \ddots \end{array}$$

- Measures the **occurrence** of symbole s in p_i
- Binary matrix
- Eg: *Vector Space Model*

Representation space

Covariance matrix

let z be the covariance matrix

$z_{ij} = (x_i - \mu)(x_j - \mu)$

Relationships between variables

Matrix $C = XX^T$ is called **contingency table** (for categorical data). It is related to the **covariance matrix** (for centered numerical data)

VARIABLES

$C =$ VARIABLES

	x^1	x^2	$\cdots$	x^N
x^1				
x^2	$\ddots$	$\vdots$		
$\vdots$	$\cdots$	$\sum_k x_k^i x_k^j$	$\cdots$	
x^N		$\vdots$	$\ddots$	

$X \left(\ i \ \right) \cdot X^T \left(\ j \ \right)$

- ⋆ symbolic : c_{ij} = number of elements with **both** symbols i and j → contingency table
- ⋆ numérique : c_{ij} measures the correlation between variables i and j → cov matrix

Representation space

Similarity table / distance matrix

Measures the distance between elements

Matrix D is called the **distance matrix**

$$X^T X = d_{ij}^2$$

$$i\begin{pmatrix} \overline{\quad x^T \quad} \end{pmatrix} \overset{j}{\begin{pmatrix} \Big| \\ x \end{pmatrix}}$$

ELEMENTS

$D =$ ELEMENTS

	p_1	p_2	$\cdots$	p_M
p_1				
P_2	$\ddots$	$\vdots$		
$\vdots$	$\cdots$	$d(\mathbf{x}_i, \mathbf{x}_j)$	$\cdots$	
p_M		$\vdots$	$\ddots$	

- ⋆ $M \times M$ matrix, can be very large
- ⋆ if d is a metric, D is p.s.d

Inertia

$P_1 \; P_2 \; \dots \; P_N$

$\rightarrow$ collum of X

$X = \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{matrix}$

- Dataset $\Omega = \{\mathbf{x}_i\}$, $i = 1 \dots, M$, $\mathbf{x} \in \mathbb{R}^N$ $\rightarrow x_N$
- there are N
- Centre of mass :

$$\mathbf{g} = \frac{1}{M} \sum_{\mathbf{x} \in \Omega} \mathbf{x}$$

- Inertia of Ω w.r.t point $\mathbf{a}$:

$$I_{\mathbf{a}} = \sum_{\mathbf{x} \in \Omega} d(\mathbf{x}, \mathbf{a})^2 \rightarrow \mathrm{var}(x) = \Sigma(x - \mu)$$

- Let $I = I_{\mathbf{g}}$ be the inertia of Ω w.r.t its center of mass $\mathbf{g}$

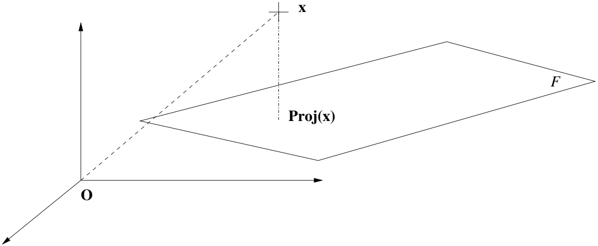
Inertia (2)

- ⋆ Inertia of Ω w.r.t subspace $\mathcal{F}$:

$$I_{\mathcal{F}} = \sum_{\mathbf{x} \in \Omega} d^2(\mathbf{x}, \mathcal{F})$$

- ⋆ $\text{Proj}_{\mathcal{F}}(\mathbf{x})$ is the orthogonal projection of $\mathbf{x}$ onto $\mathcal{F}$, then :

$$I_{\mathcal{F}} = \sum_{\mathbf{x} \in \Omega} d^2(\mathbf{x}, \text{Proj}_{\mathcal{F}}(\mathbf{x}))$$



Decomposing inertia (Huygens theorem)

- W.r.t a point:

$$\forall \, \mathbf{a} \in \mathbb{R}^N, \; I_{\mathbf{a}} = I_{\mathbf{g}} + d^2(\mathbf{a}, \mathbf{g})$$

  $\rightarrow \mathbf{g}$ point of **minimum inertia**

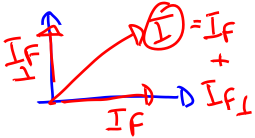
- W.r.t a subspace $\mathcal{F}$: given $\mathcal{F}_{\mathbf{g}}$ vector subspace parrallel to $\mathcal{F}$ via $\mathbf{g}$, then

$$I_{\mathcal{F}} = I_{\mathcal{F}_{\mathbf{g}}} + d^2(\mathcal{F}, \mathcal{F}_{\mathbf{g}})$$

  $\rightarrow \mathcal{F}_{\mathbf{g}}$ subspace of **minimum inertia** $//$ to $\mathcal{F}$

center of mass

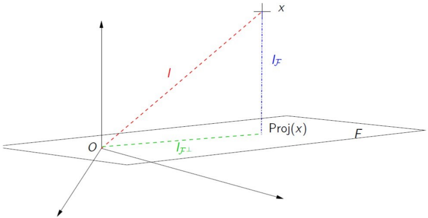
Decomposing inertia : Explained inertia



- ⋆ Centered population $\Rightarrow \mathbf{g} = O$, ($O$ origin)
- ⋆ Given $\mathcal{F}$ a vector subspace via O, then

$$I = I_{\mathcal{F}} + I_{\mathcal{F}^\perp}, \text{ avec } I_{\mathcal{F}^\perp} \text{ inertia of projected points in } \mathcal{F}^\perp$$

- ⋆ $I_{\mathcal{F}}$ is called **explained inertia** (by $\mathcal{F}$)
- ⋆ $I_{\mathcal{F}^\perp}$ is called **residual inertia** (of $\mathcal{F}$)



Matrix form

- ⋆ Centered data $\Rightarrow$

$$I = \sum_{\mathbf{x}\in\Omega} < \mathbf{x}, \mathbf{x} > \Rightarrow I = \sum_{i}^{M} \sum_{j}^{N} (x_i^j)^2 \ \Rightarrow I = \sum_{j}^{N} \sum_{i}^{M} (x_i^j)^2$$

- ⋆ if X is the matrix *variables/elements* of size $N \times M$
- ⋆ then

$$I = \text{trace}(XX^{\mathsf{T}})$$

- ⋆ matrix XX^{T} is the **inertia matrix**, or $N\times$ the **covariance matrix**

Component analysis

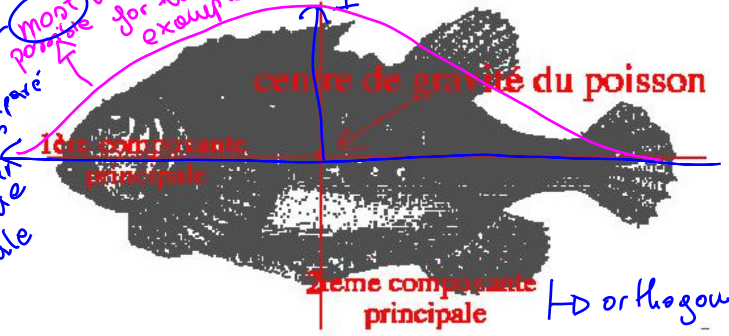
- Multivariate data analysis
  - Ω includes M points defined by N variables x^i
  - $\mathbf{x} \in \mathbb{R}^N$
- We wish to understand the spatial (resp. statistical) distribution of Ω for:
  - Data visualisation
  - Extracting the most important information
- Compression
  - Reconstruction quality
  - Easier handling of the data
- Partition against the data most important properties (features)

Principal Component Analysis (PCA)

The PCA aims at defining, for a population Ω a vector subspace within which the data is represented compactly by uncorrelated variables. These variable will be the new dimension for representing Ω.

most variance possible for this example

most comparé à d'autres choix pour composante principale



$\perp$

$\mapsto$ orthogonal

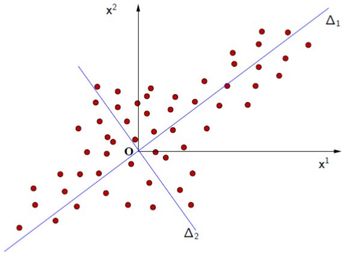
Recall on intertia

- ⋆ Given $\mathbb{R}^N = \Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_N$ the decomposition of $\mathbb{R}^N$ into 1-D orthogonal subspaces (axes Δ)
- ⋆ Given Ω, the total inertia is decomposed as

$$I = I_{\Delta_1} + I_{\Delta_2} + \cdots + I_{\Delta_N}$$

- ⋆ The PCA searches all subspaces Δ_i s.t $I_{\Delta_i} \geq I_{\Delta_{i+1}}$
- ⋆ The projections onto axes *explaining* the maximum global intertia preserve the maximum of information from the data
- ⋆ if Ω is embedded into a d-dim subspace $\Longleftrightarrow I_{\Delta_i} = 0, \ \forall i > d$

Geometric interpretation



Searching axes with maximum explained inertia

$\Longleftrightarrow$

Searching axes supporting maximal variance

Covariance matrix analysis

We search $\mathbf{u}$ minimising the quadratic error (regression)

- $\sum_i ||\mathbf{x}_i - < \mathbf{x}_i, \mathbf{u} > \mathbf{u}||^2 \simeq \sum_i \mathbf{u}^\mathsf{T} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \mathbf{u} = \mathit{Trace}(\mathbf{u}^\mathsf{T} \boldsymbol{\Sigma} \mathbf{u})$
- We add constraint $\mathbf{u}^\mathsf{T}\mathbf{u} = 1$ (to avoid collapse $||\mathbf{u}|| = 0$)

Minimisation

- Lagrange 1D, minimization of $J = \mathbf{u}^\mathsf{T} \boldsymbol{\Sigma} \mathbf{u} - \lambda(1 - \mathbf{u}\mathbf{u}) \Leftrightarrow \frac{\partial J}{\partial u} = 0 \Leftrightarrow \boldsymbol{\Sigma}\mathbf{u} - \lambda \mathbf{u} = 0$
- $\Leftrightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda \mathbf{u} \Leftrightarrow \mathbf{u}$ is an eigenvector of $\boldsymbol{\Sigma}$

Covariance matrix factorisation

Spectral decomposition

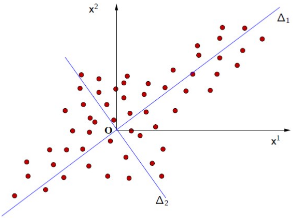
- ⋆ All dimensions are search for simultaneously
- ⋆ $\Leftrightarrow$ factorisation of Σ via eigenvalue decomposition

$$\Sigma = U \Lambda U^{\mathsf{T}}, \quad U, \Lambda \in \mathbb{R}^{N \times N}$$

- ⋆ The columns of (rotation matrix) U are the **eigenvectors** $\mathbf{u}_i$ of Σ.
- ⋆ $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_N]$ (scaling) **eigenvalues** of Σ.

Principal components

- The sorted eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ are the intertia values $I_{\Delta_1} \geq I_{\Delta_2} \geq \cdots \geq I_{\Delta_N}$
- The N sorted eigenvectors $\mathbf{u}_i$ define axes Δ_i and are called the **principal components**
- The new basis for $\mathbb{R}^N$ is now $\{\mathbf{u}_i\}_{i=1,\dots,N}$



Contribution to total inertia

$$I_{\Delta_i} = \lambda_i, \text{ et } I = \sum_i^N I_{\Delta_i}$$

Definition

1. Absolute contribution of Δ_i to I : $c_{\text{abs}}(\Delta_i/I) = \lambda_i$
2. Relative contribution: $c_{\text{rel}}(\Delta_i/I) = \frac{\lambda_i}{\lambda_1+\cdots+\lambda_N}$

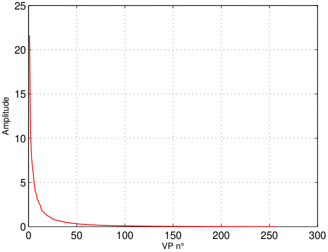
   $\Rightarrow$ Percentage of inertia explained by Δ_i
3. Percentage for the d first axes :

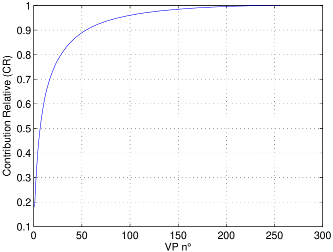
$$c_{\text{rel}}(\Delta_1 \oplus \Delta_2 \cdots \oplus \Delta_d) = \frac{\lambda_1 + \lambda_2 \cdots + \lambda_d}{\lambda_1 + \lambda_2 \cdots + \lambda_N}$$

Decreasing contribution

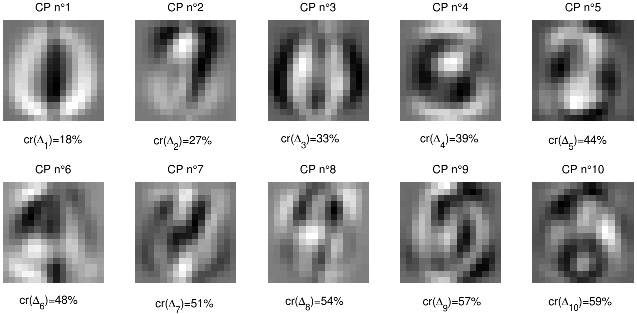
Digit data (MNIST)

7291 images 16×16 (8 bits)
$\Rightarrow \ \mathbf{x}_i \in \mathbb{R}^{256}, \ i = 1 \ldots 7291$





Explained Inertia



Projection onto principal components

The new space (basis) implies new coordinates for the data

$$y_i^j = < \mathbf{u}_j, \mathbf{x}_i >$$

$\Rightarrow$ the $j$the component of new coordinate $\mathbf{y}_i$ of a data point i is obtained by projecting $\mathbf{x}_i$ onto the jth principal component $\mathbf{u}_j$

$$\mathbf{y}_i = U^\top \mathbf{x}_i$$

Data approximation

If we retain the $d \leq N$ first components (eg $c_{\text{rel}}(\oplus_{i=1\cdots d}\Delta_i/I) \geq 90\%$). In this case, the data is approximated in the new space of reduced dimension (subspace $\oplus_{i=1\cdots d}\Delta_i$)

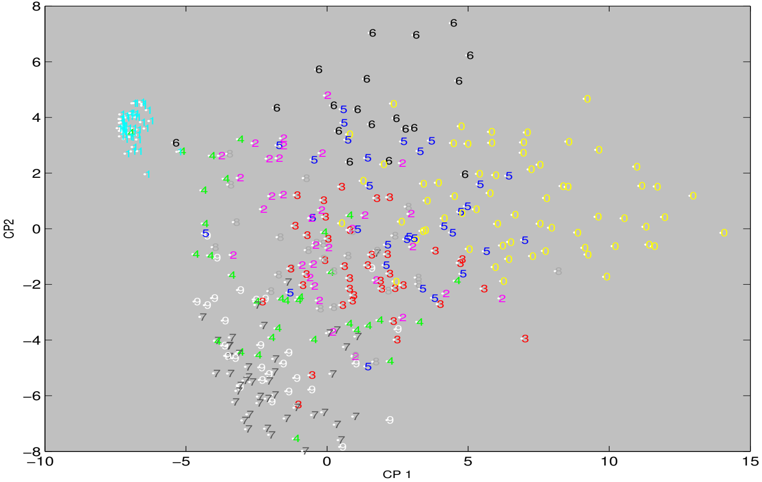
$$\tilde{\mathbf{y}}_i^2 = U_d^{\mathsf{T}}\mathbf{x}_i, \qquad \tilde{\mathbf{y}} \in \mathbb{R}^d$$

$\Rightarrow$ $U_d \in \mathbb{R}^{N\times d}$ matrix for the dth first components

- ⋆ If $d = 2$ or 3 $\rightarrow$ visualisation
- ⋆ If $d \ll N$ $\rightarrow$ compression
- ⋆ Expressivness of the d first components

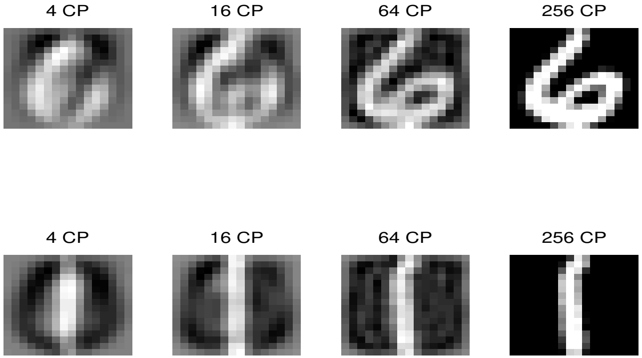
Data visualisation

$$\tilde{\mathbf{y}}_i^2 = [\mathbf{u}_1^\mathsf{T}\mathbf{x}_i, \mathbf{u}_2^\mathsf{T}\mathbf{x}_i]$$



Reconstruction

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^{d} y_i^j \mathbf{u}_j = {}^t\tilde{\mathbf{y}}_i^d U_d$$



Representation quality

- Given 2 points projected onto Δ_k
  - Far from each other on $\Delta_k \Rightarrow$, far in original space
  - Close from each other on Δ_k, no conclusion...
- We measure the representation of $\mathbf{x}_i$ on Δ_k by

$$Q_{\Delta_k}(\mathbf{x}_i) = \cos^2(\mathbf{x}_i, \mathbf{u}_k) = \frac{< \mathbf{x}_i, \mathbf{u}_k >^2}{\| \mathbf{x}_i \|^2}$$

- On subspace $E = \Delta_k \oplus \Delta_q \oplus \cdots \oplus \Delta_p$ by

$$Q_E(\mathbf{x}_i) = \cos^2(\mathbf{x}_i, \mathbf{u}_k) + \cos^2(\mathbf{x}_i, \mathbf{u}_q) + \cdots + \cos^2(\mathbf{x}_i, \mathbf{u}_p)$$

Example

Projection on main hyperplane ($\Delta_1 \oplus \Delta_2$)

Data label	0	1	2	3	4	5	6	7	8	9
Quality	0.7	**1.5**	0.4	0.2	0.7	0.2	0.5	0.9	0.4	0.8

Contribution of an element to defining axes

⋆ Absolute contribution of point i to Δ_k

$$c_{\text{abs}}(\mathbf{x}_i, \Delta_k) = \frac{1}{N} < \mathbf{x}_i, \mathbf{u}_k >^2$$

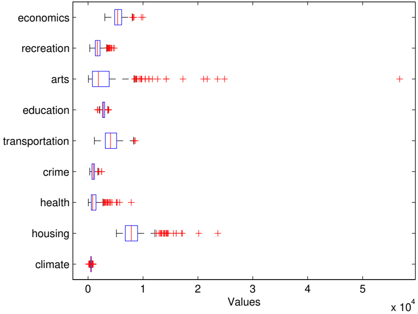
$\Longrightarrow$ The more the projection, the more the point contributes to making the axis exist

⋆ Contribution of a point to the inertia on Δ_k

$$c_{\text{rel}}(\mathbf{x}_i, \Delta_k) = \frac{\text{ca}(\mathbf{x}_i, \Delta_k)}{I_{\Delta_k}} = \frac{< \mathbf{x}_i, \mathbf{u}_k >^2}{\lambda_k}$$

PCA on scaled data

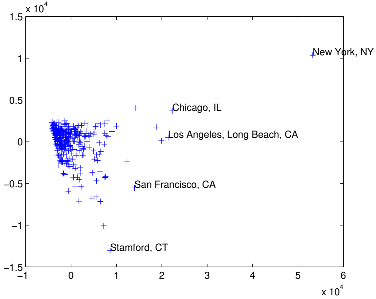
- ⋆ Heterogeneous initial variables → validity of linear combination?
- ⋆ Example : Evaluation of American cities

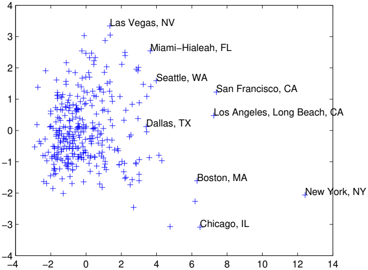


- ⋆ Scales: depend on categories 100 →> 10000
- ⋆ Necessity to scale data (by their variances)

Data scaling

- ⋆ We define metric $< .,. >_V$, with $V = diag(\sigma_1^2, \ldots, \sigma_N^2)$
- ⋆ The scaled covariance matrix Σ_V is the **correlation matrix** R of initial data
- ⋆ Spectral decomposition of R (instead of Σ)

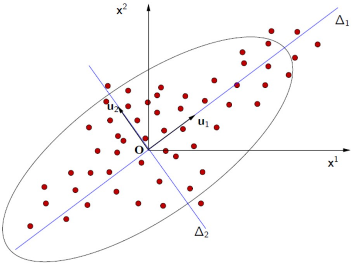




Optimality of PCA

Gaussian distribution

* PCA decomposes the covariance matrix along its eigenvalues/vectors
* Optimal basis to represent $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$



PCA is optimal to represent data whose distribution is close to Normal

Limitations

1. PCA considers the correlation between variables $\rightarrow$ linear relationships
   $\Rightarrow$ no non-linear relationship modeled
2. PCA optimises a quadratic loss
   $\Rightarrow$ sensitive to extreme values (outliers)
3. PCA is optmal for Gaussian distributed data
   $\Rightarrow$ not useful on clustered data