Data analysis : Lesson 3-4-5

Mounir Atiq

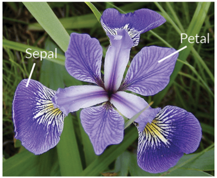
atiq.mounir@gmail.com

September 30, 2024

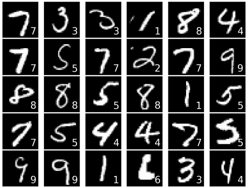
Background and Motivation

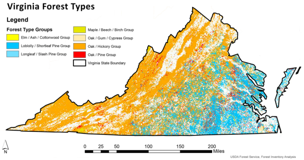
Principal Component Analysis is a data transformation tool used on continuous data for :

- Interpretation/visualization
- Dimension reduction
- First step to help clustering/classification



Iris Versicolor





- Iris dataset : 150 samples with 4 physical features, 3 classes (flower species).
- MNIST dataset : 70000 images written digits of size 28×28, 9 classes (digits).
- Forest covertype dataset : 581012 samples with 54 various environmental features, 7 classes (type of forest areas).

Projections

Definition

Let F, G be two linear subspaces of E,
F, G are **complemented subspaces** (denoted $F \bigoplus G = E$) iif
$\forall x \in E, \quad \exists!(x_F, x_G) \in F \times G : x = u + v \Longleftrightarrow F + G = E$ **and** $F \cap G = \{0_E\}$

Definition

Then, there are two applications q_F, q_G such that:
$\forall x \in E, x = q_F(x) + q_G(x)$, with $q_F(x) \in F$ and $q_G(x) \in G$

- $q_F^2 = q_F, \quad q_G^2 = q_G$
- $q_F \circ q_G = q_G \circ q_F = 0$
- $q_F + q_G = Id$
- $Im(q_F) = F = Ker(q_G), \quad Im(q_G) = G = Ker(q_F)$

q_F is the projection on F parallel to G.

Projections

Example

Consider (u, v) two vectors of $\mathbb{R}^d$ such that $\langle u, v\rangle = v^T u = 1$,

- Define $P = u^T v$, then $P^2 = P$ is a projector
- $Im(P) = Vect(u)$
- $Vect(u) \cap Vect(v)^{\perp} = \{0\}$

Then,

$$Im(P) \oplus Vect(v)^{\perp} = \mathbb{R}^d$$

So P is the projector on $Vect(u)$ parallel to $Vect(v)^{\perp}$.
In general, with $(u_1, \ldots, u_r)$ an ortho-normal family of $\mathbb{R}^d$,

- $U^T U = I_r$
- $P = UU^T$ is an orthogonal projector on $Vect(u_1, \ldots, u_r)$

Matrix decomposition

Diagonalization

Matrix $A \in \mathcal{M}_d(\mathbb{R})$ is diagonalisable iif :

$$\exists S \in \mathcal{M}_d(\mathbb{R}) : \det(S) \neq 0 \quad \text{and} \quad A = S\Lambda S^{-1}$$

,
with $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d)$ $S_{.j}$ is the j-th eigenvector associated with λ_j eigenvalue.
Moreover, if A is real symmetric, $A = S\Lambda S^T$ (Spectral Thm.)

Singular value decomposition

Consider rectangle matrix $A \in \mathcal{M}_{nd}(\mathbb{R})$,

- AA^T and A^TA are symmetric of size $n \times n$ resp. $d \times d$
- $\mathrm{rg}(AA^T) = \mathrm{rg}(A^TA) = \mathrm{rg}(A)$
- Then AA^T and A^TA share the same non-zero eigenvalues : $(\lambda_1, \ldots, \lambda_r)$

Singular value decomposition : $A = U\Lambda_r^{1/2}V^T$ with,

- $U = (u_1, \ldots, u_r)$: eigenvectors of AA^T
- $V = (v_1, \ldots, v_r)$: eigenvectors of A^TA
- $\Lambda_r^{1/2} = \mathrm{Diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$

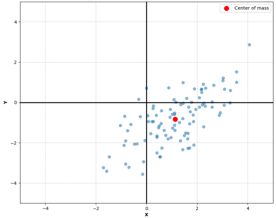
Inertia

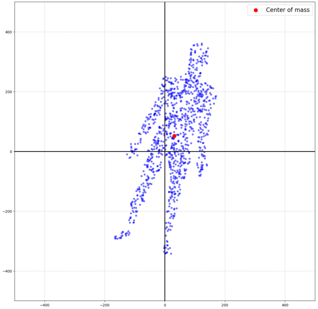
Inertia relatively to a vector

- n samples of d dimensions : $x = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$
- Sample weights : $(w_1, \ldots, w_n) \in \mathbb{R}^n : \sum_{i=1}^{n} w_i = 1$

$$I_v(x) = \sum_{i=1}^{n} w_i ||x_i - v||_2^2$$

- Inertia is minimized by the **center of mass** G.
- $I_G(x)$ quantifies the **dispersion** of x from the center of mass.





Projection inertia

Centered inertia

- n centered samples : $\widetilde{x}_i = x_i - x_G$
- Uniform weights : $w_i = \frac{1}{n}$

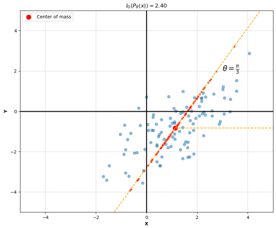
$I_G(x) = \frac{1}{n}\sum_{i=1}^{n} ||\widetilde{x}_i||_2^2 = \mathrm{Tr}(\widehat{\Sigma})$,

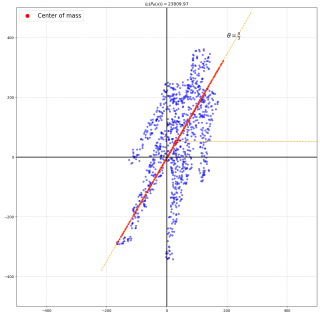
with $\widehat{\Sigma} = \frac{1}{n}\widetilde{x}^T\widetilde{x}$: empirical covariance matrix

Inertia of a projection

- Orthogonal projection on $\mathrm{Vect}(\mathrm{u}_1, \ldots, \mathrm{u}_\mathrm{r})$:
  $P_U(v) = Pv$ with $P = UU^T$
- $P^T = P$ and $P^2 = P$
- $||P_U(\widetilde{x}_i)||_2^2 = \widetilde{x}_i{}^T UU^T UU^T \widetilde{x}_i = \widetilde{x}_i{}^T P \widetilde{x}_i$

Then, $I_G\big(P_U(\widetilde{x})\big) = \mathrm{Tr}(\mathrm{P}\widehat{\Sigma}\mathrm{P}^\mathrm{T}) = \mathrm{Tr}(\mathrm{U}^\mathrm{T}\widehat{\Sigma}\mathrm{U})$





Inertia maximization

We look for a projection that maximizes the data **dispersion** to increase interpretability while looking at projected data.

Inertia maximization problem

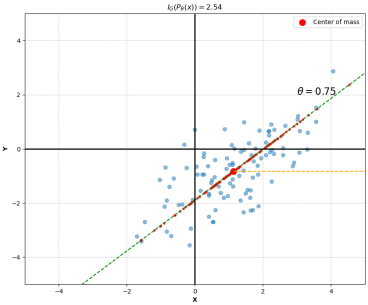
For a fixed dimension r, we look for an ortho-normal family $U = (u_1, \ldots, u_r)$ such that $I_G\big(P_U(\tilde{x})\big)$ is maximum.

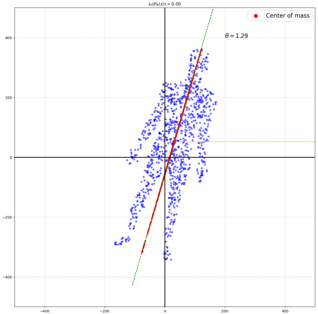
$$(u_1, \ldots, u_r) = \underset{u_k \perp u_j, ||u_k||=1}{\arg\max} \Big\{ I_G\big(P_U(\tilde{x})\big) \Big\} = \underset{||u_i||=1}{\arg\max} \Big\{ \sum_{i=1}^{n} ||P_U(\tilde{x}_i)||_2^2 \Big\}$$

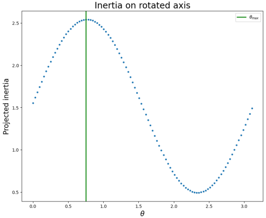
This is also the linear subspace of dimension r that approximates the best data, in the sense of mean least square error.

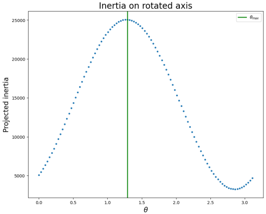
$$\iff (u_1, \ldots, u_r) = \underset{u_k \perp u_j, ||u_i||=1}{\arg\min} \Big\{ \sum_{i=1}^{n} ||\tilde{x}_i - P_U(\tilde{x}_i)||_2^2 \Big\}$$

Inertia maximization









PCA solution

Principal components subspace

The subspace of dimension r maximizing inertia is $\mathrm{Vect}(\mathrm{u}_1,\dots,\mathrm{u}_\mathrm{r})$ where u_k is the k-th eigenvector associated with the k-th biggest eigenvalue λ_k of empirical covariance matrix $\widehat{\Sigma}$.

Proof.

Firstly for $r = 1$ we search a vector u_1 such that :

$$u_1 = \underset{u\in\mathbb{R}^d, ||u||=1}{\arg\max} \left\{u^T\widehat{\Sigma}u\right\} \Longleftrightarrow (u_1, \lambda) = \underset{u\in\mathbb{R}^d, \lambda\in\mathbb{R}^d}{\arg\max} \left\{L(u,\lambda)\right\}, \quad \text{with} \quad L(u,\lambda) = u^T\widehat{\Sigma}u - \lambda(u^Tu-1)$$

$$\frac{\partial L(u,\lambda)}{\partial u} = 2(\widehat{\Sigma}u - \lambda u), \quad \text{so} \quad \frac{\partial L(u,\lambda)}{\partial u} = 0 \Longleftrightarrow \widehat{\Sigma}u = \lambda u \Longleftrightarrow u \text{ eigenvector with } \lambda \text{ eigenvalue.}$$

Then, $u^T\widehat{\Sigma}u = u^T\lambda u = \lambda||u|| = \lambda$

So u_1 eigenvector associated with maximum eigenvalue λ_{max}.

Proceed sequentially for higher dimensions, adding the constraint $\forall k \in \{1,\dots,j\}, u_{j+1} \perp u_k$. □

PCA solution

Principal components computation

X a random variable of dimension d,
Principal vectors are the columns of U and principal components are new random variables obtained by the transformation $Y = U^T(X - \mathbb{E}[X])$.

- $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_2)$
- $\Sigma = \mathrm{Cov(X)} = \mathrm{U\Lambda U^T}$
- $\mathbb{E}[Y_i] = 0, \mathrm{Var(Y_i)} = \lambda_\mathrm{i}$
- $i \neq j \implies \mathrm{Cov(Y_i, Yj)} = 0$

Given n samples $x = (x_1, \ldots, x_n)$ of dimension d (as a matrix of size $n \times d$),

- $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^d$, so for each column j, $\widetilde{x}_j = x_j - \widehat{\mu}_j$
- For each i-th data : $y_i = U^T\widetilde{x}_i$ so samples of **principal components** are : $y = \widetilde{x}U$
- New empirical covariance matrix $\widehat{\Sigma}_{YY} = \Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_\mathrm{d})$

Data pre-processing for PCA

Scaling and weighting depend on the data and the application.

Weighting/scaling

- Weighting samples by $(w_1, \dots, w_n)$: $\tilde{x} = W^{1/2}x$, with $W^{1/2} = \mathrm{Diag}(\sqrt{\mathrm{w_1}}, \dots, \sqrt{\mathrm{w_n}})$
- Scaling variables : $\tilde{x} = xM^{1/2}$, with M symmetric definite-positive
- Usual scaling : $M^{1/2} = \mathrm{Diag}(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_\mathrm{d}})$, then $\frac{1}{n}\tilde{x}^T\tilde{x}$ is the correlation matrix.

Then PCA can be seen as Singular Value Decomposition applied on $xM^{1/2}$.

This corresponds to replace $||x||_2^2$ by $||x||_M^2 = x^T Mx$ in inertia maximization problem.
Most of the time, we just take $D = \frac{1}{n} I_n$ and $M = I_d$.

Data pre-processing for PCA

PCA steps

1. Pre-processing of data :
   - Always center data
   - Depending on situation, scale/weight data
2. PCA algorithm
   - Compute empirical covariance matrix $\widehat{\Sigma}$
   - Compute eigenvector/eigenvalue of $\widehat{\Sigma}$ and sort them
   - Project data on principal vectors
3. Interpretation
   - Individuals : contributions and representation quality
   - Variables : explained variance ratio and correlation circle
   - Dimension reduction : restrict data to $r < d$ principal components

Data representation on the first factorial plane

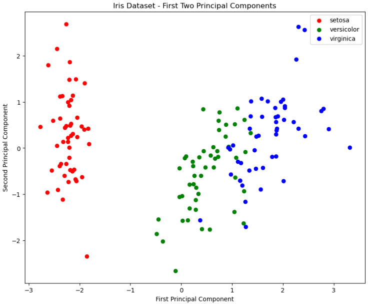


Figure 1: 2-d principal components representation of Iris dataset

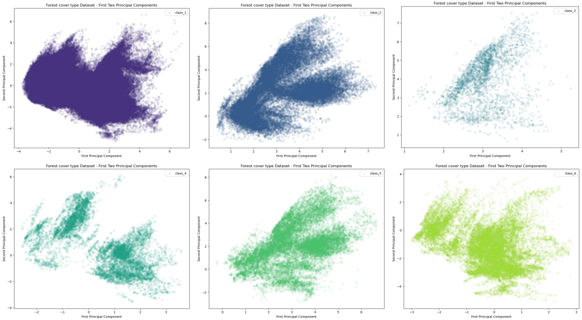


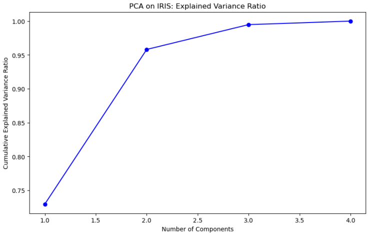
Figure 2: 2-d principal components representation of each class of Forest Covertype dataset

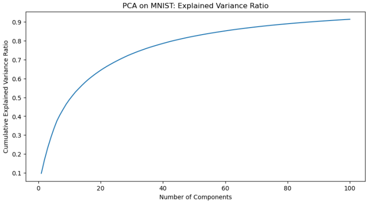
Explained variance/inertia ratio

Each components carries a certain proportion of the total empirical variance.

Principal components computation

- Total inertia : $I_G(x) = \sum_{k=1}^{d} I_G\big(P_{u_k}(x)\big) = \sum_{k=1}^{d} \lambda_k$
- Variance ratio explained by u_i : $r_i = \frac{\lambda_i}{\sum_{k=1}^{d} \lambda_k}$
- Variance ratio explained by the first q components : $R_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{d} \lambda_k}$





Individuals contribution

In the same way we can consider the variance contribution by one or several samples.

Principal components computation

- k-th component of sample j : $c_{jk} = \langle \tilde{x}_j, u_k \rangle$
- Total variance of the k-th component : $\mathrm{Var}(C_k) = I_G\left(P_{u_k}(x)\right) = \lambda_k = \frac{1}{n}\sum_{j=1}^{n} c_{jk}^2$
- Contribution j-th individual on k-th component : $\gamma_{jk} = \frac{1}{n}\frac{c_{jk}^2}{\lambda_k}$
- Total contribution of j-th individual : $\gamma_j = \frac{1}{n}\frac{\sum_{k=1}^{d} c_{jk}^2}{\sum_{k=1}^{d} \lambda_k}$

Individual contributions can help to detect outliers and sometimes aberrant samples that drag too much one or several components towards it. They may impact too much representation of the rest of data and it might be better to exclude such samples.

Quality of data representation

The representation quality of a sample $\tilde{x}_j$ by the axis u_k is how much the sample is *aligned* with u_k.

Representation quality of an individual

- Quality of k-th component on j-th individual : $q_{jk} = \cos^2(\tilde{\mathrm{x}}_\mathrm{j}, \mathrm{u}_\mathrm{k}) = \frac{\langle \tilde{\mathrm{x}}_\mathrm{j}, \mathrm{u}_\mathrm{k} \rangle^2}{||\tilde{\mathrm{x}}_\mathrm{j}||_2^2} = \frac{\mathrm{c}_\mathrm{jk}^2}{\sum_{\mathrm{k}=1}^{\mathrm{d}} \mathrm{c}_\mathrm{jk}^2}$
- Representation quality of r principal components : $\sum_{k=1}^{r} q_{jk} = \frac{\sum_{k=1}^{r} c_{jk}^2}{\sum_{k=1}^{d} c_{jk}^2}$

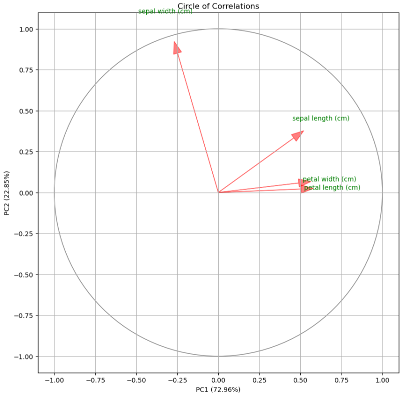
Correlation circle

By inverting orthogonal matrix U, we can express initial variables X_j in function of principal components C_k. This is the inverse PCA transformation.

Inverse PCA transformation

- k-th coordinate of X_j : $\frac{\langle X_j, C_k\rangle}{||C_k||} = \frac{\langle \sum_{l=1}^{d} u_{jl} C_l, C_k\rangle}{||C_k||} = u_{jk}||C_k|| = \sqrt{\lambda_k} u_{jk}$
- Correlation coefficient : $\rho_{jk} = \sqrt{\frac{\lambda_k}{\mathrm{Var(X_j)}}} u_{jk}$
- Partial reconstruction of X_j (with $r < d$) : $X_j \approx \sum_{k=1}^{r} \frac{\langle X_j, C_k\rangle}{||C_k||} C_k$

If data are **scaled** correlations between variables and principal components are directly these coordinates (which are contained in a sphere of radius 1).
But sometimes it is preferable to not scale data.



Latent factors model

Assuming centered variables can be reconstructed with r components : $\widetilde{X} = \sum_{k=1}^{r} c_k U_k + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \sigma I_r)$

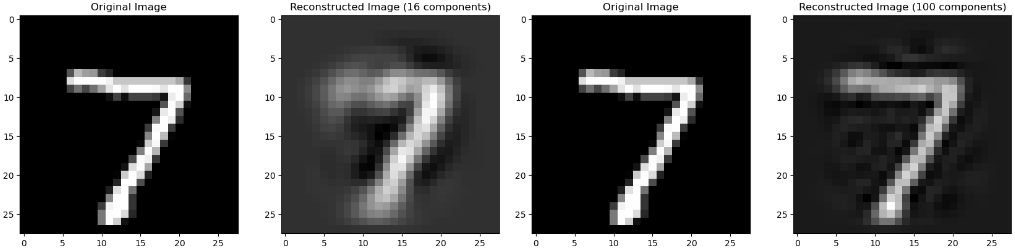


Figure 3: Partially reconstructed images for $r = 16$ and $r = 100$

1. Introduction
2. Reminder : linear algebra
3. Principal component analysis
4. Canonical correlation analysis
   - Introduction
   - Correlation maximization problem
   - Interpretation
   - Associated hypothesis testing
   - CCA steps summary
5. Factorial correspondence analysis

Motivation

Canonical correlation analysis purpose is to study relations between different groups of variables and to point out their proximity.
A common situation is to try to explain one group of variables $Y = (Y_1, \dots, Y_q)$ by another group $X = (X_1, \dots, X_p)$.
Application examples :

- Genetics : phenotype features on one side, genotype features on the other.
- Economics : price/demand on one side, product characteristic on the other.
- Design : multi-objective performance variables on one side, design aspects on the other.

Comparison with PCA :

- PCA : search of a new ortho-normal basis $(u_1, \dots, u_d)$.
  For a given dimension $r < d$, the projection on $\mathrm{Vect(u_1, \dots, u_r)}$ maximizes the **inertia**.
- CCA : joint search two ortho-normal basis $(u_1, \dots, u_p)$ and $(v_1, \dots, v_q)$.
  One for each group of variable $X = (X_1, \dots, X_p)$ and $X = (Y_1, \dots, Y_q)$.
  For a given dimension $r < \min(p, q)$ the correlation between the projections $\mathrm{Vect(u_1, \dots, u_r)}$ and $\mathrm{Vect(v_1, \dots, v_r)}$ is maximized.

Example : 3d Gaussian

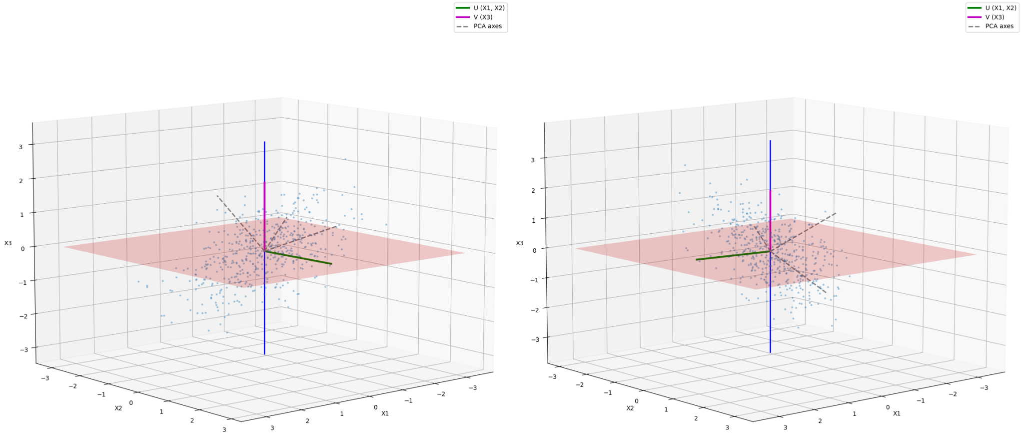


Figure 4: PCA and CCA axes on 3d Gaussian example

Example : 3d Gaussian

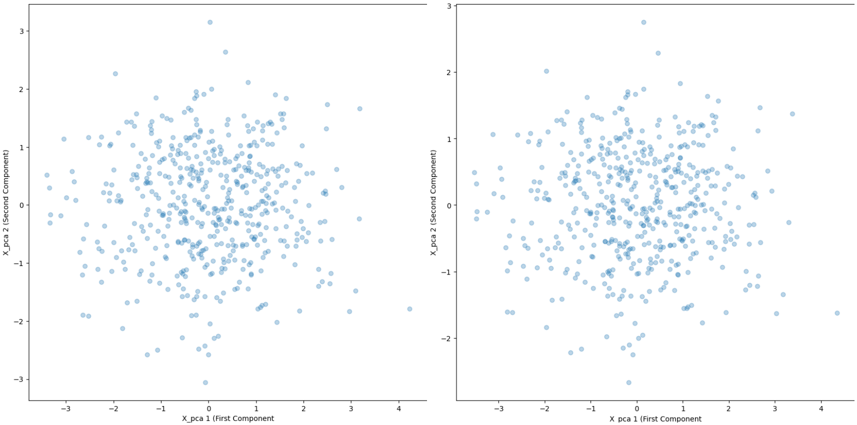


Figure 5: Projection on first PCA axes

Example : 3d Gaussian

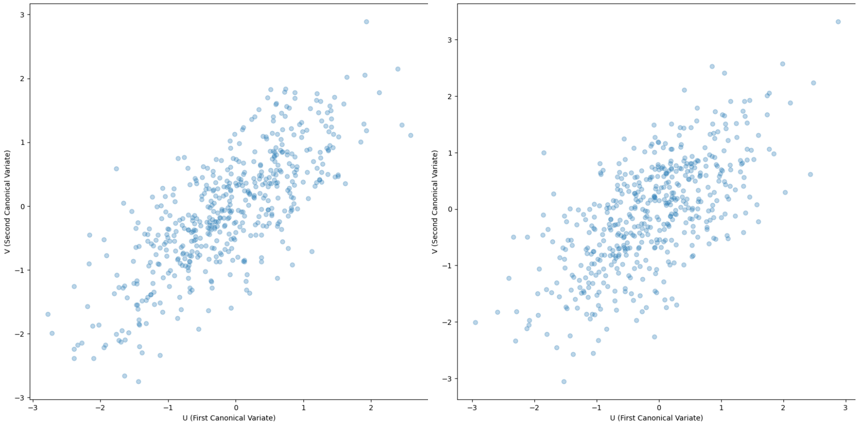


Figure 6: Projection on first CCA axes

Correlation maximization

Correlation of linear combinations

Consider two groups of random variables : $X = (X_1, \dots, X_p)$ and $X = (Y_1, \dots, Y_q)$,

The objective is to find $a \in \mathbb{R}^p, b \in \mathbb{R}^q$ maximizing correlation $\rho_{V,W}$ with $V = a^T X = \sum_{k=1}^{p} a_k X_k$ and $W = b^T Y = \sum_{k=1}^{q} b_k Y_k$.

- $\rho_{V,W} = \frac{\mathrm{Cov(V,W)}}{\sqrt{\mathrm{Var(V)Var(W)}}}$ and $\forall c \in \mathbb{R}, \rho_{cV,W} = \rho_{V,cW} = \rho_{V,W}$
- $\mathrm{Cov(V,W)} = \sum_{k=1}^{p} \sum_{j=1}^{q} a_k b_j \mathrm{Cov}(X_k, Y_j) = a^T \Sigma_{XY} b$
- $\mathrm{Var(V)} = a^T \Sigma_{XX} a$ and $\mathrm{Var(W)} = b^T \Sigma_{YY} b$

$$\boxed{(a^*, b^*) = \underset{(a,b)}{\arg\max}\{a^T \Sigma_{XY} b\} \quad \text{subject to} \quad a^T \Sigma_{XX} a = 1 \text{ and } b^T \Sigma_{YY} b = 1} \tag{1}$$

Correlation maximization

Solution to correlation maximization problem

Denote $k = \text{rank}(\Sigma_{XY})$, $\quad \forall 1 \leq k \leq r$,
Solution of $\underset{(a_k, b_k)}{\arg\max}\{a_k^T \Sigma_{XY} b_k\}$ subject to :

- $a_k^T \Sigma_{XX} a_k = 1 \quad$ and $\quad b_k^T \Sigma_{XX} b_k = 1$
- $\forall 1 \leq j \leq k-1, a_k^T \Sigma_{XX} a_j = 0 \quad$ and $\quad b_k^T \Sigma_{XX} b_j = 0$

is given by :

- Maximum equals $\sqrt{\lambda_k}$ the k-th singular value of $M = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$
- a_k is the associated k-th eigenvector of $\left(\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^T\right)$
- b_k is the associated k-th eigenvector of $\left(\Sigma_{YY}^{-1} \Sigma_{XY}^T \Sigma_{XX}^{-1} \Sigma_{XY}\right)$

Correlation maximization

Proof.

Define Lagrangian : $\mathscr{L}(a,b,\lambda_x,\lambda_y) = a^T\Sigma_{XY}b - \lambda_x\left(a^T\Sigma_{XX}a - 1\right) - \lambda_y\left(b^T\Sigma_{YY}b - 1\right)$

- $\frac{\partial\mathscr{L}(a,b,\lambda_x,\lambda_y)}{\partial a} = \Sigma_{XY}b - 2\lambda_x\Sigma_{XX}a$
- $\frac{\partial\mathscr{L}(a,b,\lambda_x,\lambda_y)}{\partial b} = \Sigma_{XY}^T a - 2\lambda_y\Sigma_{YY}b$

$$\begin{cases}\frac{\partial\mathscr{L}(a,b,\lambda_x,\lambda_y)}{\partial a} = 0\\ \frac{\partial\mathscr{L}(a,b,\lambda_x,\lambda_y)}{\partial b} = 0\end{cases} \qquad \begin{cases}2\lambda_x a = \Sigma_{XX}^{-1}\Sigma_{XY}b\\ 2\lambda_y b = \Sigma_{YY}^{-1}\Sigma_{XY}^T a\end{cases} \qquad \begin{cases}2\lambda_x a = \frac{1}{2\lambda_y}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T a\\ 2\lambda_y b = \frac{1}{2\lambda_x}\Sigma_{YY}^{-1}\Sigma_{XY}^T\Sigma_{XX}^{-1}\Sigma_{XY}b\end{cases}$$

Thus,

- a is eigenvector of $\left(\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T\right)$
- b is eigenvector of $\left(\Sigma_{YY}^{-1}\Sigma_{XY}^T\Sigma_{XX}^{-1}\Sigma_{XY}\right)$

Moreover, we can rewrite maximization depending only on a :
Maximizing $a^T\Sigma_{XY}b \implies a^T\Sigma_{XY}b = \frac{1}{2\lambda_y}a^T\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T a$
and $a^T\Sigma_{XY}b = 2\lambda_y$

We would like to get from (a,b) new vectors γ,δ and find a matrix M such that γ is eigenvector of MM^T and δ is eigenvector of M^TM. (To apply Singular Value Decomposition)

□

Correlation maximization

Proof.

By denoting :

- $M = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$
- $\gamma = \Sigma_{XX}^{1/2} a$
- $\delta = \Sigma_{YY}^{1/2} b$
- $MM^T = \Sigma_{XX}^{-1/2}\left(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T\right)\Sigma_{XX}^{-1/2}$
- $M^TM = \Sigma_{YY}^{-1/2}\left(\Sigma_{XY}^T\Sigma_{XX}^{-1}\Sigma_{XY}\right)\Sigma_{YY}^{-1/2}$

Equation (1) is equivalent to :

$$\underset{||\gamma||=1}{\arg\max}\{\gamma^T MM^T\gamma\} \quad \text{and} \quad \underset{||\delta||=1}{\arg\max}\{\delta^T M^T M\delta\}$$

Then, eigenvectors $(\gamma_1,\dots,\gamma_r)$ of MM^T and eigenvectors $(\delta_1,\dots,\delta_r)$ of M^TM share the same non-zero eigenvalues $\lambda_1 \geq \dots \geq \lambda_r$ and λ_1 maximizes this quantity (with $r = \text{rank}(\Sigma_{XY})$. Thus, $(a^*, b^*) = (\Sigma_{XX}^{-1/2}\gamma_1, \Sigma_{YY}^{-1/2}\delta_1)$ and

$\max\{a^T\Sigma_{XY}b\} = \sqrt{\lambda_1}$.

Similarly, for further dimension $(k \leq r)$: $(a_k^*, b_k^*) = (\Sigma_{XX}^{-1/2}\gamma_k, \Sigma_{YY}^{-1/2}\delta_k)$ and $\max\{a^T\Sigma_{XY}b\} = \sqrt{\lambda_k}$ $\square$

Canonical variables

New variable representation

The initial variables are $X = (X_1, \dots, X_p)$ and $Y = (Y_1, \dots, Y_q)$ following :

$$\mathbb{E}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \end{pmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{XX} \end{pmatrix}$$

After CCA we get new variable $V = (V_1, \dots, V_r)$ and $W = (W_1, \dots, W_r)$, defined as $V_k = a_k^T X$ and $W_k = b_k^T Y$.

$$\mathbb{E}\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} a_1^T \mathbb{E}[X] \\ \dots \\ a_r^T \mathbb{E}[X] \\ b_1^T \mathbb{E}[Y] \\ \dots \\ b_r^T \mathbb{E}[Y] \end{pmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} I_r & \Lambda^{1/2} \\ \Lambda^{1/2} & I_r \end{pmatrix} \quad \text{with} \quad \Lambda^{1/2} = \mathrm{Diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_\mathrm{r}})$$

Canonical variables

Canonical variable computation

Denote $E_x = \mathrm{Vect}(\tilde{\mathrm{x}}_{.1}, \ldots, \tilde{\mathrm{x}}_{.\mathrm{p}})$ and $E_y = \mathrm{Vect}(\tilde{\mathrm{y}}_{.1}, \ldots, \tilde{\mathrm{y}}_{.\mathrm{q}})$ linear subspaces of $\mathbb{R}^n$, and P_x and P_y orthogonal projections on E_x and E_y.
v_k and w_k represent n samples of dimension 1,

- $v_k = \tilde{x} a_k \quad$ and $\quad w_k = \tilde{y} b_k$
- $P_x = \tilde{x}(\tilde{x}^T\tilde{x})^{-1}\tilde{x}^T \quad$ and $\quad P_y = \tilde{y}(\tilde{y}^T\tilde{y})^{-1}\tilde{y}^T$

Then,

- v_k is eigenvector of $P_x P_y$ associated with λ_k
- w_k is eigenvector of $P_y P_x$ associated with λ_k

Canonical correlations / Example : physiological-exercise DB

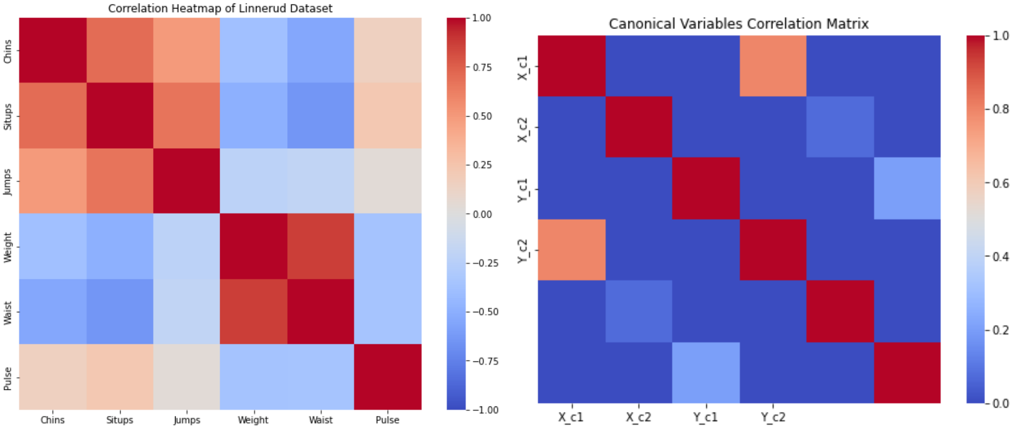
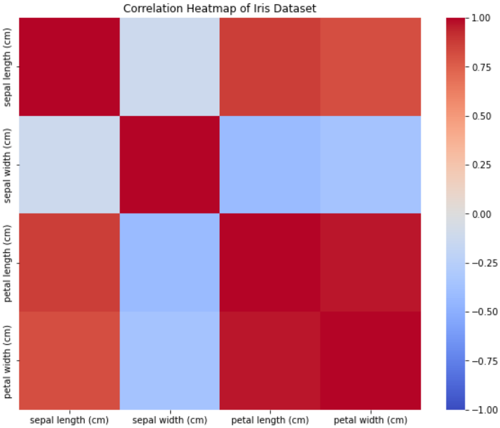


Figure 7: Initial correlation matrix and canonical correlation matrix

Canonical correlations / Example : Iris DB



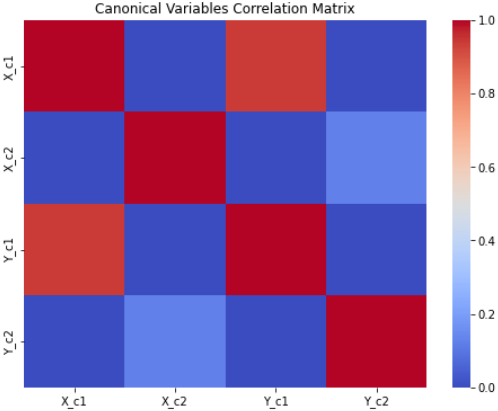


Figure 8: Initial correlation matrix and canonical correlation matrix

Canonical variables / Example : physiological-exercise DB

	Ex. 1	Ex. 2	Ex. 3
Chins	0.73	0.64	0.24
Situps	0.82	−0.05	0.57
Jumps	0.16	0.23	0.96

	Physio. 1	Physio. 2	Physio. 3
Weight	−0.62	0.13	−0.77
Waist	−0.92	0.03	0.38
Pulse	0.33	−0.94	0.04

Table 1: Comparison between canonical variables and corresponding initial variables

	Physio. 1	Physio. 2	Physio. 3
Chins	0.58	0.05	0.05
Situps	0.65	0	0.11
Jumps	0.13	0.02	0.19

	Ex. 1	Ex. 2	Ex. 3
Weight	−0.49	0.01	−0.15
Waist	−0.74	0.	−0.08
Pulse	0.26	−0.06	0.01

Table 2: Cross comparison between canonical variables and initial variables of the other group

Canonical variables / Example : Iris DB

	Sepal 1	Sepal 2
Sepal length (cm)	0.93	0.37
Sepal length (cm)	−0.48	0.88

	Petal 1	Petal 2
Petal length (cm)	0.99	0.14
Petal width (cm)	0.91	0.40

Table 3: Comparison between canonical variables and corresponding initial variables

	Petal 1	Petal 2
Sepal length (cm)	0.87	0.05
Sepal width (cm)	−0.45	0.11

	Sepal 1	Sepal 2
Petal length (cm)	0.93	0.02
Petal width (cm)	0.86	0.05

Table 4: Cross comparison between canonical variables and initial variables of the other group

Correlation circle / Example : physiological-exercise DB

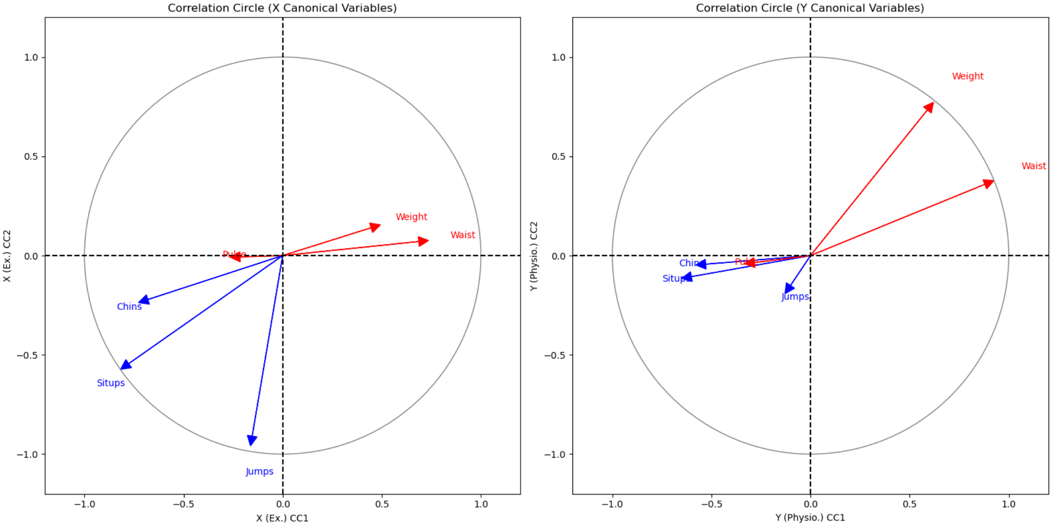


Figure 9: Correlation circles on two first canonical variables of each group

Correlation circle / Example : Iris DB

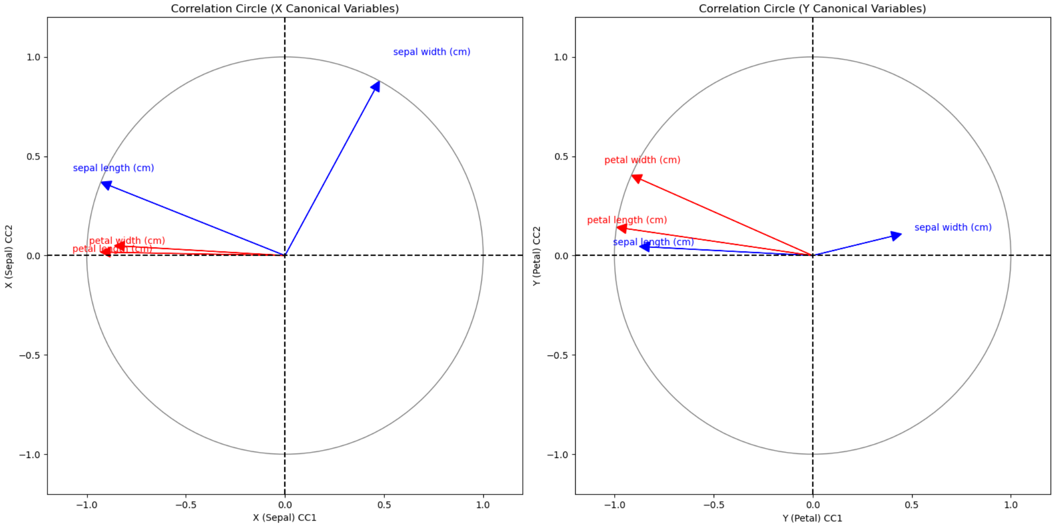


Figure 10: Correlation circles on two first canonical variables of each group

Data projection

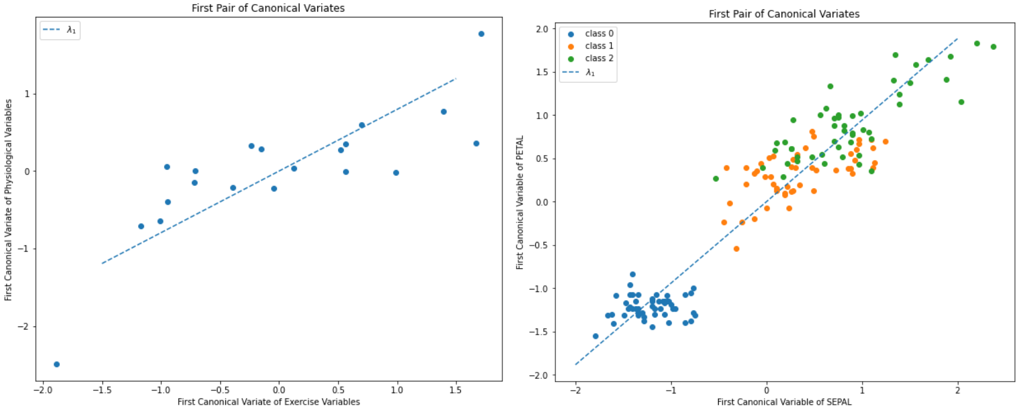


Figure 11: Data 2d representation using the first canonical variable of each group

Hypothesis testing

For the CCA to bring usefull information, we need the two variable groups to be correlated in some way.

Non-correlation hypothesis testing

$\mathscr{H}_0$: "non-correlation of canonical variables",

Testing the significance of first k canonical correlations :

$$\xi_n = -(n - \frac{p+q+3}{2}) \log \prod_{i=1}^{k} (1-\lambda_i) \sim \chi^2_{pq}$$

Testing the significance of canonical correlations from $l+1$ to k :

$$\xi_n = -(n - \frac{p+q+3}{2}) \log \prod_{i=l+1}^{k} (1-\lambda_i) \sim \chi^2_{(p-l)(q-l)}$$

Hypothesis testing / Examples

By computing this statistics and comparing with chi-square quantile Tables, we can interpret the significance of canonical correlations.

All canonical correlations :

- Physiological-exercise DB : $n = 20$, $p = 3$, $q = 3$ / $\xi_n = 29.25$
- χ^2_9 quantile of order 99% : 21.67
- Iris DB : $n = 150$, $p = 2$, $q = 2$ / $\xi_n = 433.93$
- χ^2_4 quantile of order 99% : 13.28

All canonical correlations except the first one :

- Physiological-exercise DB : $\xi_n = 4.64$
- χ^2_4 quantile of order 99% : 13.28
- Iris DB : $\xi_n = 19.38$
- χ^2_1 quantile of order 99% : 6.63

CCA algorithm summary

CCA steps

1. Pre-processing of data :
   - Always center both data groups x and y.
2. CCA algorithm computations
   - Compute empirical covariance matrices $\widehat{\Sigma}_{XX}, \widehat{\Sigma}_{YY}$ and $\widehat{\Sigma}_{XY}$.
   - Apply singular value decomposition of $M = \widehat{\Sigma}_{XX}^{-1/2} \widehat{\Sigma}_{XY} \widehat{\Sigma}_{YY}^{-1/2}$.
   - Sort singular values $(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$
   - As well as associated left/right eigenvectors $(\gamma_1, \ldots, \gamma_r)$ and $(\delta_1, \ldots, \delta_r)$.
   - Deduce group X CCA axes : $a_k = \widehat{\Sigma}_{XX}^{-1/2} \gamma_k$ and group Y CCA axes $b_k = \widehat{\Sigma}_{YY}^{-1/2} \delta_k$.
3. Interpretation
   - Canonical variables : $V_k = a_k^T X$ and $W_k = b_k^T Y$
   - Canonical correlations : $\sqrt{\lambda_k}$
   - Ensure correlation between two groups : hypothesis testing.
   - Data representation : represent data in each (V_k, W_k) plane.
   - Dimension reduction : restrict data to $k < r$ canonical variables.
   - Prediction : use canonical variables to built linear model between the two groups.

Motivation

Factorial correspondence analysis objective is to study, represent and synthesize information contained on joint distribution of categorical variables from the perspective of their inter-dependencies.

But the quantities PCA and CCA algorithms maximize are **not well defined on categorical variables**.
It is first designed on a couple of two categorical variables $X \in \{1, \ldots p\}$ and $Y \in \{1, \ldots q\}$ and uses discrete probability estimation to :

- Define a notion of center of mass to the joint distribution.
- Define a notion of dispersion of the joint distribution from this center.
- Use this new dispersion measure to apply PCA like method.

Then, **multiple correspondence analysis** is generalization to a wider group of several categorical variables.

Example : national assembly representatives previous jobs

	DEM	DR	ECOS	EPR	GDR	HOR	LFI-NFP	LIOT	NI	RN	SOC	UDR
Agriculteur	2	1	1	2	0	0	0	0	0	3	2	0
Artisants/Commerçants/Chef entr.	3	4	1	14	0	5	1	2	0	7	1	4
Sans activité	0	1	0	0	0	0	0	0	0	1	0	0
Cadres sup.	26	36	29	65	9	23	42	19	6	74	52	10
Employés	1	1	0	2	3	1	10	0	0	14	1	0
Ouvriers	0	0	0	0	0	0	3	0	0	3	0	0
Professions intermédiaires	2	0	3	7	5	3	11	0	0	14	8	0
Retraités	1	3	2	6	0	1	1	0	1	5	2	0
Non declarés	1	1	2	1	0	0	4	0	0	5	0	2

Table 5: Contingency table of French national assembly representatives, by political group and professional background
Source : https://datan.fr/statistiques/deputes-origine-sociale

Hypothesis testing

Independence hypothesis testing

Consider n couples of iid discrete random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ in $\{1, \dots, p\}$ and $\{1, \dots, q\}$,

- $C_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k, Y_k = i,j\}}$: number of samples in each cell of contingency table.
- $C_{i.} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k = i\}}$, $C_{.j} = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k = j\}}$: total number of samples for each possible value of X (resp. Y).
- $E_{ij} = \frac{C_{i.} C_{.j}}{n}$: expected value C_{ij} under independence.

$$d^2_{\chi^2}(C) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(C_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{\mathscr{L}} \chi^2_{(p-1)(q-1)} \qquad \text{(under independence)}$$

It is called **chi-square distance** associated with contingency table C.

Row / column profiles

Discrete probabilities estimators

- $p_{ij} = \frac{C_{ij}}{n}$ is an estimator of $\mathbb{P}(X = i, Y = j)$
- $p_{i\cdot} = \sum_{j=1}^{q} p_{ij} \left(\text{resp. } p_{\cdot j} = \sum_{i=1}^{p} p_{ij} \right)$ is an estimator of $\mathbb{P}(X = i) \left(\text{resp. } \mathbb{P}(Y = j) \right)$
- $\frac{p_{ij}}{p_{i\cdot}} \left(\text{resp. } \frac{p_{ij}}{p_{\cdot j}} \right)$ is an estimator of $\mathbb{P}(Y = j / X = i) \left(\text{resp. } \mathbb{P}(X = i / Y = j) \right)$

To get rid of the influence of the number of samples n we define :

Row / column profiles

- Average row (resp. column) profiles : $\{p_{1\cdot}, \dots, p_{p\cdot}\}$(resp. $\{p_{\cdot 1}, \dots, p_{\cdot q}\}$)
- Profile of i-th row : $(\frac{p_{i1}}{p_{i\cdot}}, \dots, \frac{p_{iq}}{p_{i\cdot}})$
- Profile of j-th column : $(\frac{p_{1j}}{p_{\cdot j}}, \dots, \frac{p_{pj}}{p_{\cdot j}})$

Exercice

Show that : $$\frac{(p_{ij} - p_{i\cdot}.p_{\cdot j})^2}{p_{i\cdot}.p_{\cdot j}} = \frac{1}{n} \frac{(C_{ij} - E_{ij})^2}{E_{ij}}$$

Row / column profiles

	DEM	DR	ECOS	EPR	GDR	HOR	LFI-NFP	LIOT	NI	RN	SOC	UDR
Agriculteur	0.18	0.09	0.09	0.18	0.00	0.00	0.00	0.00	0.00	0.27	0.18	0.00
Artisants/Commerçants/Chef entr.	0.07	0.10	0.02	0.33	0.00	0.12	0.02	0.05	0.00	0.17	0.02	0.10
Sans activité	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00
Cadres sup.	0.07	0.09	0.07	0.17	0.02	0.06	0.11	0.05	0.02	0.19	0.13	0.03
Employés	0.03	0.03	0.00	0.06	0.09	0.03	0.30	0.00	0.00	0.42	0.03	0.00
Ouvriers	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00
Professions intermédiaires	0.04	0.00	0.06	0.13	0.09	0.06	0.21	0.00	0.00	0.26	0.15	0.00
Retraités	0.05	0.14	0.09	0.27	0.00	0.05	0.05	0.00	0.05	0.23	0.09	0.00
Non declarés	0.06	0.06	0.12	0.06	0.00	0.00	0.25	0.00	0.00	0.31	0.00	0.12

Table 6: Row profiles assembly representatives dataset
(Summing along columns equals 1)

	DEM	DR	ECOS	EPR	GDR	HOR	LFI-NFP	LIOT	NI	RN	SOC	UDR
Agriculteur	0.06	0.02	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.00
Artisants/Commerçants/Chef entr.	0.08	0.09	0.03	0.14	0.00	0.15	0.01	0.10	0.00	0.06	0.02	0.25
Sans activité	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Cadres sup.	0.72	0.77	0.76	0.67	0.53	0.70	0.58	0.90	0.86	0.59	0.79	0.63
Employés	0.03	0.02	0.00	0.02	0.18	0.03	0.14	0.00	0.00	0.11	0.02	0.00
Ouvriers	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00
Professions intermédiaires	0.06	0.00	0.08	0.07	0.29	0.09	0.15	0.00	0.00	0.11	0.12	0.00
Retraités	0.03	0.06	0.05	0.06	0.00	0.03	0.01	0.00	0.14	0.04	0.03	0.00
Non declarés	0.03	0.02	0.05	0.01	0.00	0.00	0.06	0.00	0.00	0.04	0.00	0.12

Table 7: Column profiles of assembly representatives dataset
(Summing along rows equals 1)

Link with inertia

χ^2 distance can be interpreted as a kind of inertia on contingency table with specific weighting and scaling.

Inertia of row/column profiles

Considering row profiles as $\mathbb{R}^q$ vectors:

- Each row i represents estimators of $\mathbb{P}(Y/X=i)$ conditional probabilities relatively to i : $z_i = \frac{p_{ij}}{p_{i\cdot}}$
- Weighting of i-th sample : $w_i = p_{i\cdot}$
- Center of mass of rows : $G_{row} = (p_{\cdot 1}, \ldots, p_{\cdot q})$
- Scaling : $z_i^{(s)} = \frac{1}{\sqrt{p_{\cdot j}}} \frac{p_{ij}}{p_{i\cdot}}$

$$||z_i - G_{row}||_2^2 = \sum_{j=1}^{q} (\frac{p_{ij}}{p_{i\cdot}} - p_{\cdot j})^2 \quad \text{and} \quad ||z_i^{(s)} - G_{row}^{(s)}||_2^2 = \sum_{j=1}^{q} \frac{1}{p_{\cdot j}} (\frac{p_{ij}}{p_{i\cdot}} - p_{\cdot j})^2$$

$$I_G(z^{(s)}) = \sum_{i=1}^{p} w_i ||z_i^{(s)} - G_{row}^{(s)}||_2^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{p_{i\cdot}}{p_{\cdot j}} \left(\frac{p_{ij}}{p_{i\cdot}} - p_{\cdot j} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}} = \frac{1}{n} d_{\chi^2}^2$$

Same result for column profiles by inverting row/column roles.

χ^2 distance as a measure of dispersion

χ^2 distance quantifies how far observed samples are from **independence** situation.

The purpose of AFC is to emphasize **relations** between the two categorical variables.

Then, the approach is to transform either row profiles or column profiles to **maximize χ^2 distance**.

Standardized residuals

Residuals are the equivalent to centered-scaled data on contingency tables,

- $R_{ij} = \frac{C_{ij} - E_{ij}}{\sqrt{E_{ij}}}$
- $d^2_{\chi^2} = \sum_{i=1}^{p} \sum_{j=1}^{q} R_{ij}^2 = Tr(R^T R) = Tr(RR^T)$

Example

Politic groups / professional background :

- χ^2 distance of contingency table : 132.427
- χ^2_{88} quantile of order 99% : 121.767

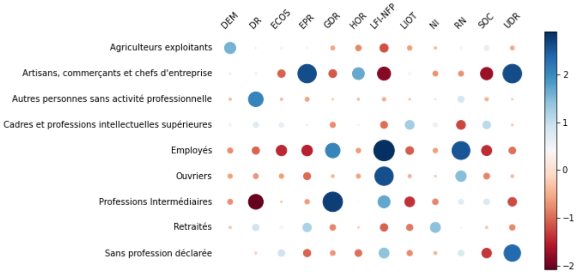


Figure 12: Residuals represent positive (blue) / negative (red) associations between values of categorical variables.

FCA is a double PCA on residuals

Inertia of row/column profiles

Then maximizing χ^2 distance is equivalent to inertia maximization on standardized residuals R.

This means we can apply **double PCA** : considering R as a data matrix for row profiles and R^T for column profiles.

- **Centered data** : $C - E$
- **Weighting** : $D_L^{-1/2}(C-E)$, with $D_L = \mathrm{Diag}(\mathrm{C}_{1\cdot}, \ldots, \mathrm{C}_{\mathrm{p}\cdot})$
- **Scaling** : $R = D_L^{-1/2}(C-E)D_C^{-1/2}$, with $D_C = \mathrm{Diag}(\mathrm{C}_{\cdot 1}, \ldots, \mathrm{C}_{\cdot \mathrm{q}})$
- Moreover $E = \frac{1}{n} D_L 1_p 1_q D_C$

(Transpose C and E and invert roles of D_L and D_C for column profiles.)

Thus we perform Singular Value Decomposition on matrix R :

- $R = V\Lambda^{1/2}U^T$, with $\Lambda^{1/2} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_\mathrm{r})$
- U is composed by ortho-normed eigenvectors of $R^T R$, rotation matrix of the PCA of R.
- V is composed by ortho-normed eigenvectors of RR^T, rotation matrix of the PCA of R^T.
- $Ru_k = \sqrt{\lambda_k} v_k \in \mathbb{R}^p$: k-th principal components of row profiles.
- $R^T v_k = \sqrt{\lambda_k} u_k \in \mathbb{R}^q$: k-th principal components of column profiles.

Visual representation

As for the PCA, we project original weighted data on eigenvectors $(u_1, \dots, u_r)$ for row profiles and on eigenvectors $(v_1, \dots, v_r)$ for column profiles.

New coordinates are called **principal factors**:

- Rows principal factors :
  $f_k^{(row)} = D_L^{-1/2} R u_k = \sqrt{\lambda_k} D_L^{-1/2} v_k$
- Columns principal factors :
  $f_k^{(col)} = D_C^{-1/2} R^T v_k = \sqrt{\lambda_k} D_C^{-1/2} u_k$
- Closeness of two row points (resp. column points) means their conditional distribution are close.
- Closeness of a row point (resp. column point) to the origin means it is close to average row (resp. column) profile
- Closeness between a row point and a column point means this row plays an important role in this column.

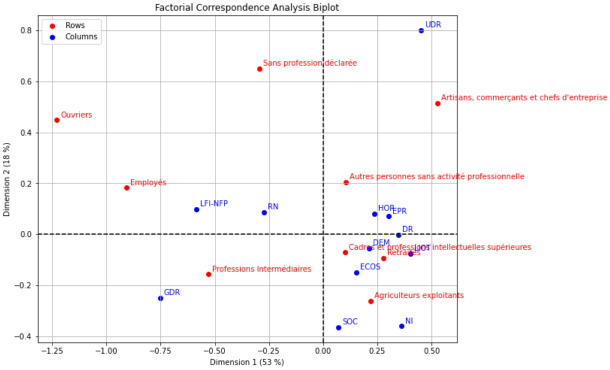


Figure 13: Row/columns projection on first two principal factors

Reconstruction of contingency table

Reconstruction formula

- By definition $C_{ij} = E_{ij}\left(1 + \frac{R_{ij}}{\sqrt{E_{ij}}}\right)$
- By Singular Value Decomposition : $R_{ij} = \left(V\Lambda^{1/2}U^T\right)_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k} V_{ik} U_{jk}, \quad$ with $r = \min(p-1, q-1)$

Then,

$$C_{ij} = E_{ij}\left(1 + \frac{\sum_{k=1}^{r} \sqrt{\lambda_k} V_{ik} U_{jk}}{\sqrt{E_{ij}}}\right)$$

So the FCA can be seen as a decomposition of observed samples deviation from independence situation into r components of decreasing importance according to eigenvalues.

In the same way as previous methods, we can project our data (here row/column profiles) on a subspace of lower dimension.

Choice of projection subspace dimension

In the FCA, the notion of inertia is linked with χ^2 distribution.

- Total inertia follows asymptotically a χ^2 distribution : $\sum_{k=1}^{r} \lambda_k \sim \chi^2_{(p-1)(q-1)}$
- In the same way : $\sum_{k=1+d}^{r} \lambda_k \sim \chi^2_{(p-d-1)(q-d-1)}$

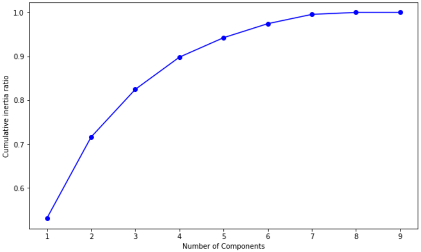


Figure 14: χ^2 distance ratio by number of factors on politic representatives dataset

Dimension reduction decision

So a strategy to fix this dimension is :

- Choose a quantile order α.
- Search the lowest d such that $\sum_{k=1+d}^{r} \lambda_k$ is inferior to the α-quantile of $\chi^2_{(p-d-1)(q-d-1)}$.
- Select d as the reduced dimension subspace to project data.

Contributions and representation quality

Similarly to the PCA we can define notions of contribution and representation quality (both on row/column profiles):

- Quality of k-th factor on j-th row profile : $\dfrac{{f_{jk}^{(row)}}^2}{\sum\limits_{k=1}^{d} {f_{jk}^{(row)}}^2}$
- Contribution of j-th row profile to the k-th factor : $\dfrac{C_i.{f_{jk}^{(row)}}^2}{\lambda_k}$

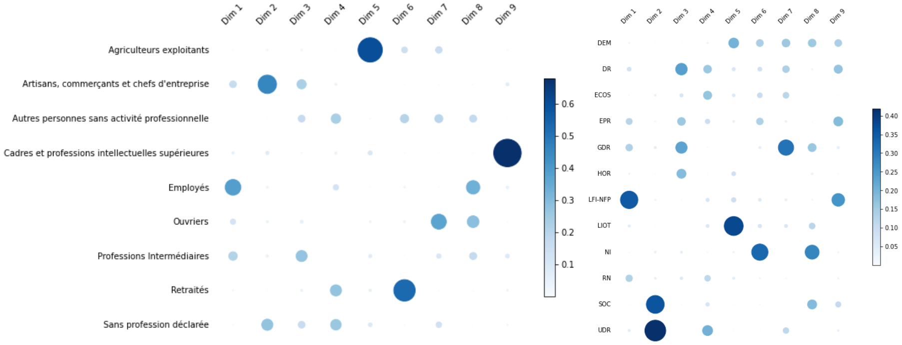


Figure 15: Row/columns contributions on principal factors

Example of FCA on student data

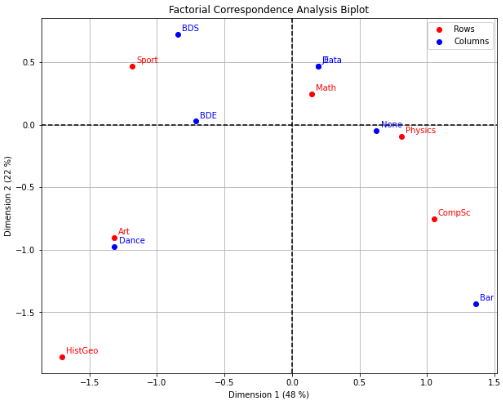


Figure 16: Row/column projection on first two principal factors

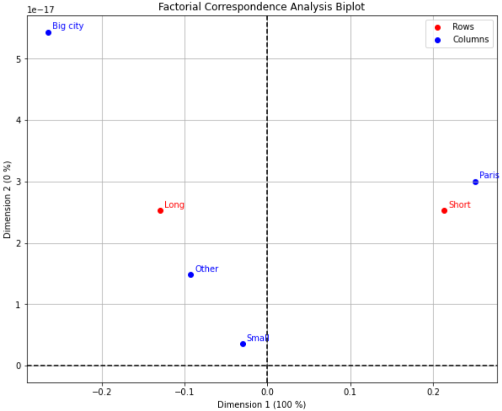


Figure 17: Row/column projection on first two principal factors

Guttman effect

When the two categorical variables are redundant, we observe a typical shape on the first two components projection.

	col1	col2	col3	col4	col5
row1	15	60	5	0	0
row2	4	200	25	3	0
row3	2	82	100	25	1
row4	1	5	40	90	8
row5	0	3	7	20	10

It happens in particular while discrete but still quantitative variables are considered.

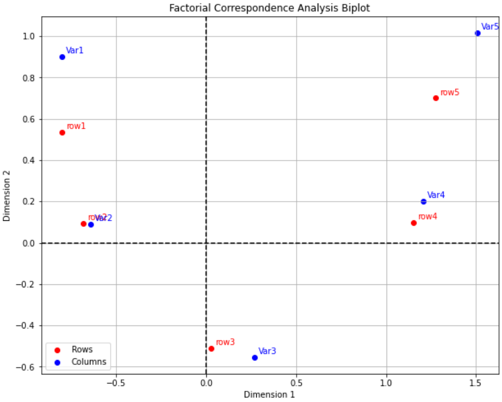
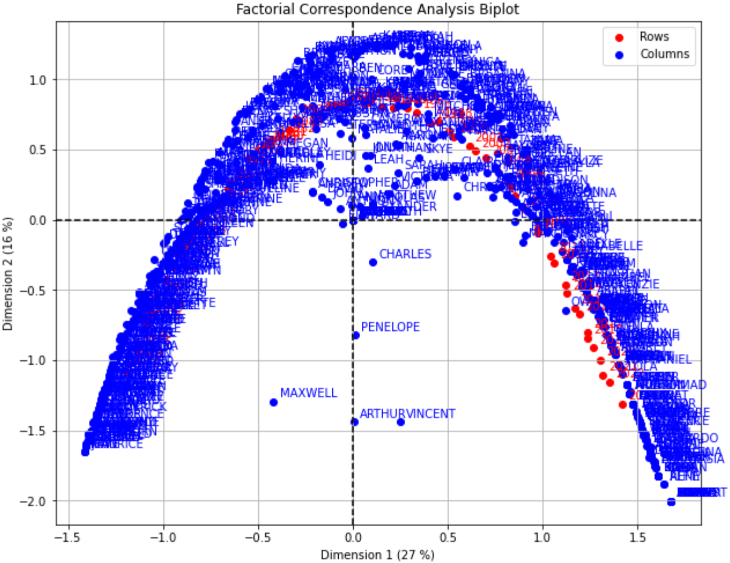


Figure 18: Row/column projection with Guttman effect

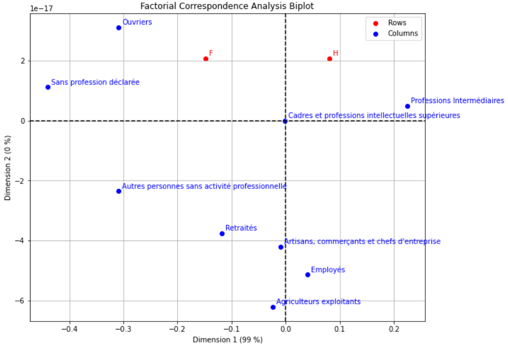
Guttman effect

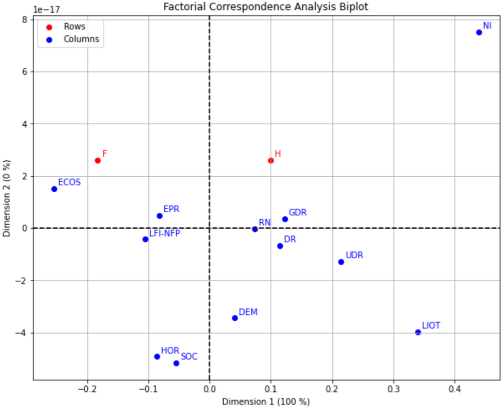


Generalization to numerous categorical variables

What if we want to add gender to the study of politic representatives ?

One approach is to consider each contingency table for each couple of variables.





Burt table

P1	P2	P3	P4	P5	P6	P7	P8	P9	DEM	DR	ECO	EPR	GDR	HOR	LFFP	LIOT	NI	RN	SOC	UDR	Fem.	Male
11	0	0	0	0	0	0	0	0	2	1	1	2	0	0	0	0	0	3	2	0	4	7
0	42	0	0	0	0	0	0	0	3	4	1	14	0	5	1	2	0	7	1	4	15	27
0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1
0	0	0	391	0	0	0	0	0	26	36	29	65	9	23	42	19	6	74	52	10	138	253
0	0	0	0	33	0	0	0	0	1	1	0	2	3	1	10	0	0	14	1	0	11	22
0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	3	0	0	3	0	0	3	3
0	0	0	0	0	0	53	0	0	2	0	3	7	5	3	11	0	0	14	8	0	13	40
0	0	0	0	0	0	0	22	0	1	3	2	6	0	1	1	0	1	5	2	0	9	13
0	0	0	0	0	0	0	0	16	1	1	2	1	0	0	4	0	0	5	0	2	9	7
2	3	0	26	1	0	2	1	1	36	0	0	0	0	0	0	0	0	0	0	0	12	24
1	4	1	36	1	0	0	3	1	0	47	0	0	0	0	0	0	0	0	0	0	14	33
1	1	0	29	0	0	3	2	2	0	0	38	0	0	0	0	0	0	0	0	0	18	20
2	14	0	65	2	0	7	6	1	0	0	0	97	0	0	0	0	0	0	0	0	38	59
0	0	0	9	3	0	5	0	0	0	0	0	0	17	0	0	0	0	0	0	0	5	12
0	5	0	23	1	0	3	1	0	0	0	0	0	0	33	0	0	0	0	0	0	13	20
0	1	0	42	10	3	11	1	4	0	0	0	0	0	0	72	0	0	0	0	0	29	43
0	2	0	19	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	4	17
0	0	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	7	0	0	0	1	6
3	7	1	74	14	3	14	5	5	0	0	0	0	0	0	0	0	0	126	0	0	40	86
2	1	0	52	1	0	8	2	0	0	0	0	0	0	0	0	0	0	0	66	0	25	41
0	4	0	10	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	16	4	12
4	15	1	138	11	3	13	9	9	12	14	18	38	5	13	29	4	1	40	25	4	203	0
7	27	1	253	22	3	40	13	7	24	33	20	59	12	20	43	17	6	86	41	12	0	373

Table 8: Burt Table of 3 categorical variables on politics dataset

- Each diagonal block is a diagonal matrix representing the total for each possible value of each variable.
- Other blocks represent contingency matrices between each couple of categorical variables.

FCA on Burt table

Applying FCA to Burt Table allows to represent all variables on the same subspace.

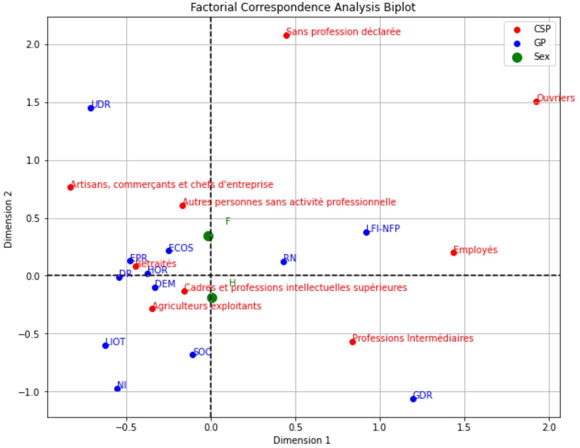


Figure 20: Example of FCA projection using Burt Table (on politic representatives dataset)