Introduction to statistical learning

2.1 Unsupervised learning: Principal Component Analysis

V. Lefieux

June 2018

Table of contents

Table of contents

Cars

Technical information on 18 cars:

- car model (MOD),
- cylinder capacity (CYL),
- power (POW),
- length (LEN),
- width (WID).
- weight (WGT),
- speed (SPD),
- finish (FIN),
- price (PRI).

Source: (Saporta, 2011)

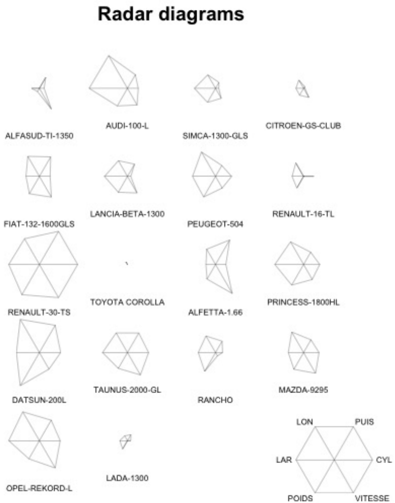
Data

MOD	CYL	POW	LEN	WID	WGT	SPD	FIN	PRI
ALFASUD-TI-1350	1350	79	393	161	870	165	B	30570
AUDI-100-L	1588	85	468	177	1110	160	TB	39990
SIMCA-1300-GLS	1294	68	424	168	1050	152	M	29600
CITROEN-GS-CLUB	1222	59	412	161	930	151	M	28250
FIAT-132-1600GLS	1585	98	439	164	1105	165	B	34900
LANCIA-BETA-1300	1297	82	429	169	1080	160	TB	35480
PEUGEOT-504	1796	79	449	169	1160	154	B	32300
RENAULT-16-TL	1565	55	424	163	1010	140	B	32000
RENAULT-30-TS	2664	128	452	173	1320	180	TB	47700
TOYOTA COROLLA	1166	55	399	157	815	140	M	26540
ALFETTA-1.66	1570	109	428	162	1060	175	TB	42395
PRINCESS-1800HL	1798	82	445	172	1160	158	B	33990
DATSUN-200L	1998	115	469	169	1370	160	TB	43980
TAUNUS-2000-GL	1993	98	438	170	1080	167	B	35010
RANCHO	1442	80	431	166	1129	144	TB	39450
MAZDA-9295	1769	83	440	165	1095	165	M	27900
OPEL-REKORD-L	1979	100	459	173	1120	173	B	32700
LADA-1300	1294	68	404	161	955	140	M	22100

Radar diagrams I

Given the small number of variables here, one can represent each individual with a radar diagram.

Radar diagrams II

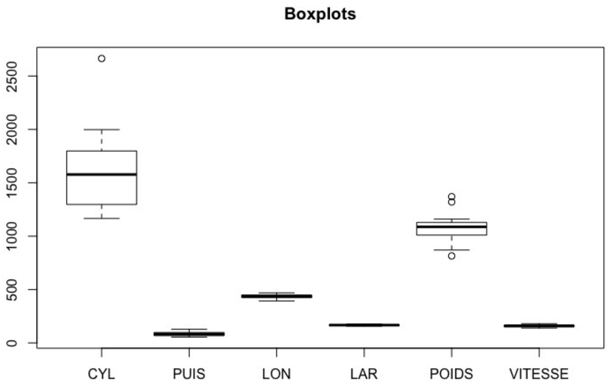


Radar diagrams III

There are some radars, small or big, but most of the time harmonious: variables have the same evolution.

It's possible to distinguish some models with a specific shape, for example small sport cars (faster compared to other small cars).

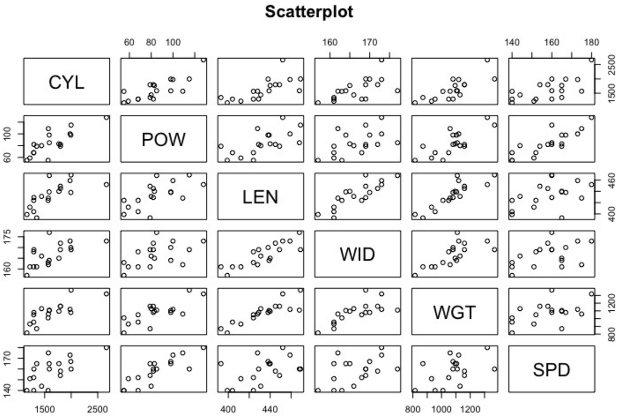
Boxplots


Boxplots
0
500
1000
1500
2000
2500
CYL
PUIS
LON
LAR
POIDS
VITESSE

Standard deviations

CYL	POW	LEN	WID	WGT	SPD
373.9	20.4	22.1	5.3	137.0	12.1

Scatter plots



Correlation matrix

	CYL	POW	LEN	WID	WGT	SPD
CYL	1.0000000	0.7966277	0.7014619	0.6297572	0.7889520	0.6649340
POW	0.7966277	1.0000000	0.6413624	0.5208320	0.7652930	0.8443795
LEN	0.7014619	0.6413624	1.0000000	0.8492664	0.8680903	0.4759285
WID	0.6297572	0.5208320	0.8492664	1.0000000	0.7168739	0.4729453
WGT	0.7889520	0.7652930	0.8680903	0.7168739	1.0000000	0.4775956
SPD	0.6649340	0.8443795	0.4759285	0.4729453	0.4775956	1.0000000

Table of contents

Data

p quantitative variables measured on n individuals.

The data set that is represented in terms of an $n \times p$ matrix:

$$\mathbb{X} = \left(x_i^j\right)_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}} ,$$

where the n rows are the individuals and the p columns are the variables.

x_i^j: value of X^j measured on individual i.

Dataset matrix

		Variables				
		1	...	j	...	p
Individuals	1	x_1^1	...	x_1^j	...	x_1^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	x_i^1	...	x_i^j	...	x_i^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	n	x_n^1	...	x_n^j	...	x_n^p

Individuals and variables

Commonly individual i refers to vector:

$$X_i = \left(x_i^1, \ldots, x_i^p\right)^\top$$

and variable j to vector:

$$X^j = \left(x_1^j, \ldots, x_n^j\right)^\top .$$

Weights

The sample should be representative: a miniature of the population it comes from. If not, one assign to each individual i a weight ω_i (e.g from a survey design):

- $\forall i \in \{1, \ldots, n\} : \omega_i > 0$,
- $\sum_{i=1}^{n} \omega_i = 1$.

One consider the matrix:

$$W = \operatorname{diag}(\omega_1, \ldots, \omega_n) \ .$$

Usually weights are uniform:

$$\forall i \in \{1, \ldots, n\} : \omega_i = \frac{1}{n} \ ,$$

that is:

$$W = \frac{1}{n} \mathrm{I}_n \ .$$

Barycenter

The barycenter of the data set is:

$$G = \mathbb{X}^\top W \, \mathbf{1}_n = \sum_{i=1}^{n} \omega_i X_i$$

where $\mathbf{1}_n$ is a n dimensional vector with all its components equal to 1.

Problem

To study and interpret $\mathbb{X}$, we would like to plot the n individuals in the p-dimensional space $\mathbb{R}^p$. Obviously it's impossible and we need to reduce the number of variables.

Choosing some variables among the data set would be totally arbitrary.

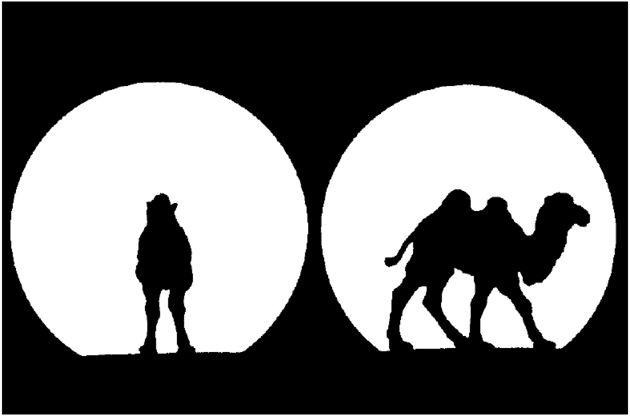
Principal Component Analysis (PCA) uses an orthogonal linear transformation to convert a set of correlated variables into a set of linearly uncorrelated variables (principal components).

The goal of PCA is to summarize the correlations among the data set with a smaller set of variables: the data set can often be interpreted in just a few principal components.

PCA steps

- Calculate the principal components which iteratively extracts the maximum variance from the data.
- Determine how many principal components should be considered.
- Interpret the principal components.
- Analyse the individuals projections onto principal components (in practice 2-3).

Cow, camel or horse ?



Reference: (Fénelon, 2000)

Table of contents

Standardisation I

Consider the sample mean and standard deviation:

$$\overline{x^j} = \sum_{i=1}^{n} \omega_i \, x_i^j \ ,$$

$$s_j^2 = \sum_{i=1}^{n} \omega_i \left(x_i^j - \overline{x^j} \right)^2 \ .$$

The centered representation of the individual i is:

$$\forall j \in \{1, \ldots, p\} : y_i^j = x_i^j - \overline{x^j} \ .$$

The standardized representation of the individual i is:

$$\forall j \in \{1, \ldots, p\} : z_i^j = \frac{x_i^j - \overline{x^j}}{s_j} \ .$$

Standardisation II

PCA capture the total variance in the data set, PCA results depend on the scales of variables.
So PCA requires that the input variables have similar scales of measurement.

- PCA: based on the centered representation.
- Standardized PCA: based on the standardized representation.

Dissimilarity metric

We consider the following distance between 2 individuals i_1 and i_2 in $\mathbb{R}^p$:

$$d_M^2\left(i_1, i_2\right) = \left(X_{1\,i} - X_{i_2}\right)^\top Q\left(X_{i_1} - X_{i_2}\right)$$

with:

- $M = \mathrm{I}_p$ for a PCA,
- $M = \mathrm{diag}\left(\frac{1}{s_1^2}, \ldots, \frac{1}{s_p^2}\right) := D_{\frac{1}{s^2}}$ for a standardized PCA.

For $(x, y) \in \mathbb{R}^p \times \mathbb{R}^p$, we define the inner product:

$$\langle x, y \rangle_M = x^\top M\, y$$

and the norm:

$$\|x\|_M^2 = x^\top M\, x\ .$$

Inertia I

The total inertia $\mathcal{I}_{tot}$ of the data set is:

$$\mathcal{I}_{tot} = \sum_{i=1}^{n} \omega_i \, d_M^2 \left(i, G \right) \, .$$

Inertia II

The projected inertia $\mathcal{I}_H$ of the data set on the affine subspace H is:

$$\mathcal{I}_H = \sum_{i=1}^{n} \omega_i \, d_M^2\Big(P_H(i), P_H(G)\Big) \ .$$

where P_H is the orthogonal projection on H.

$\mathcal{I}_H$ is a measure of the remaining information after projection on H. The aim is to find H for which $\mathcal{I}_H$ is maximized.

Inertia III

The residual inertia $\mathcal{J}_H$ of the data set is:

$$\mathcal{J}_H = \sum_{i=1}^{n} \omega_i \, d_M^2 \left(i, P_H(i)\right)$$

Inerties IV

It can be shown (Huygens theorem) that:

$$\mathcal{I}_{tot} = \mathcal{I}_H + \mathcal{J}_H \ .$$

Moreover the subspace H contains G, so $P_H(G) = G$, and:

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \omega_i \, \omega_{i'} \, d_M^2 \left(i, i'\right) = 2 \, \mathcal{I}_{tot} \ ,$$

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \omega_i \, \omega_{i'} \, d_M^2 \left(P_H(i), P_H\left(i'\right) \right) = 2 \, \mathcal{I}_H \ .$$

Conclusion for inertia

In conclusion, we search a subspace H that:

- maximizes $\mathcal{I}_H$,
- minimizes $\mathcal{J}_H$,
- maximizes the sum of the distances between the projected individuals on H.

Variance-covariance and correlation matrixes

Let S be the variance-covariance matrix of the data set:

$$S = \mathbb{X}^\top W \, \mathbb{X} - G^\top G \ .$$

For $(i,j) \in \{1, \ldots, p\}^2$, the element (j_1, j_2) of the matrix is:

$$s_{i,j} = \text{cov}\left(X^i, X^j\right)$$

The i-th diagonal element is s_i^2.

The correlation matrix R is:

$$R = D_{\frac{1}{s}} \, S \, D_{\frac{1}{s}} = \mathbb{Z}^\top W \, \mathbb{Z} \ .$$

Note that cov (respectively var, corr) is the empirical covariance (respectively variance, correlation).

Total inertia

It can be shown that:

$$\mathcal{I}_{tot} = \mathrm{Tr}\,(MS) \ .$$

So:

- For a PCA:

$$\mathcal{I}_{tot} = \sum_{j=1}^{p} s_j^2 \ .$$

- For a standardized PCA:

$$\mathcal{I}_{tot} = p \ .$$

Eigenvalues and eigenvectors I

The matrix SM is:

- Symmetric
  So SM is diagonalizable, there exists an orthogonal matrix P ($PP^\top = \mathrm{I}_p$) a diagonal matrix which entries are eigenvalues $(\lambda_1, \ldots, \lambda_p)$ such that:

  $$SM = P \ \mathrm{diag}\left(\lambda_1, \ldots, \lambda_p\right) \ P^\top \ .$$

- Positive semidefinite
  Eigenvalues are nonnegative.

Eigenvalues and eigenvectors II

We consider that eigenvalues $(\lambda_1, \ldots, \lambda_p)$ are in descending order:

$$\lambda_1 \geq \ldots \geq \lambda_p \geq 0 \ .$$

For $\alpha \in \{1, \ldots, p\}$, we consider eigenvector u_α associated to λ_α, such that $\|u_\alpha\|_{M^{-1}} = 1$.

This vector is called factor.

Idea

Finding the k-dimensional subspace which maximizes the projected inertia is equivalent to find the k eigenvectors associated to the k biggest eigenvalues of the matrix SM.

Theorem

For a k-dimensional space:

- H_k which maximizes projected inertia is:

$$H_k = \text{vect}\left(u_1, \ldots, u_k\right) \ .$$

- Projected inertia on α-th factor u_α is equal to the α-th eigenvalue:

$$I_{u_\alpha} = \lambda_\alpha \ .$$

  Each components eigenvalue represents how much variance it explains.

- Projected inertia on H_k is the sum of the k biggest eigenvalues:

$$I_{H_k} = \sum_{\alpha=1}^{k} \lambda_\alpha \ .$$

The resolution of the problem can be iteratively computed.

Principal components

We define p new variables called principal components.

For $\alpha \in \{1, \ldots, p\}$, principal component is:

$$C^{\alpha} = \sum_{j=1}^{p} u_{\alpha}^{j} X^{j} = \mathbb{X} u_{\alpha} \in \mathbb{R}^{n} \ ,$$

Principal components are uncorrelated:

$$\forall (\alpha, \beta) \in \{1, \ldots, p\}^2 : \mathrm{cov}\left(C^{\alpha}, C^{\beta}\right) = \begin{cases} 0 & \text{si } \alpha \neq \beta \\ \lambda_{\alpha} & \text{si } \alpha = \beta \end{cases} \ .$$

Other interpretation

- The first principal component is the linear combination of the variables that has the maximal variance among all linear combinations.
- The second principal component is the linear combination of the variables that has the maximal variance among all linear combinations uncorrelated to the first principal component
- . . .

Correlation between the principal components and the original variables I

Case study
Problem
Method
PCA on the case study
References

For the PCA, correlation between α-th principal component and j-th original variables is:

$$\text{corr}\left(C^{\alpha}, X^{j}\right) = \frac{\sqrt{\lambda_{\alpha}}}{s_j} u_{\alpha}^{j} \ .$$

So:

$$\sum_{j=1}^{p} s_j^2 \, \text{corr}^2\left(C^{\alpha}, X^{j}\right) = \lambda_{\alpha} \ .$$

Correlation between the principal components and the original variables II

For the standardized PCA, correlation between α-th principal component and j-th original variables is:

$$\operatorname{corr}\left(C^{\alpha}, X^{j}\right) = \sqrt{\lambda_{\alpha}}\; u_{\alpha}^{j} \; .$$

So:

$$\sum_{j=1}^{p} \operatorname{corr}^{2}\left(C^{\alpha}, X^{j}\right) = \lambda_{\alpha} \; .$$

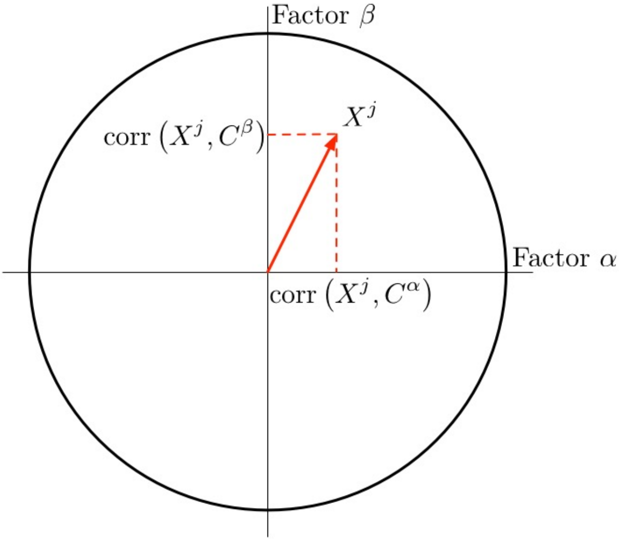
Correlation circle I

It's possible to visualize correlations between principal components ans original variables.

In the factorial space (α, β), we plot the vector X^j with coordinates $\left(\text{corr}\left(C^\alpha, X^j\right), \text{corr}\left(C^\beta, X^j\right)\right)$.

In the case of the standardized PCA, vectors are inside the unit circle called correlation circle.

Correlation circle II



Correlation circle III

In the factorial space (α, β):

- A variable close to the correlation circle can be considered well represented by the factorial space.
- 2 variables close to the correlation circle, nearly orthogonal, have a small correlation.

Determination of the number of principal components

- Kayser criterium
  Retain components with eigenvalues greater than their mean (1 in standardized PCA).

- Scree plot criterium
  Find in the scree plot a steep curve followed by a bend and then a flat or horizontal line (retain as number of principal components the last point before the flat line).

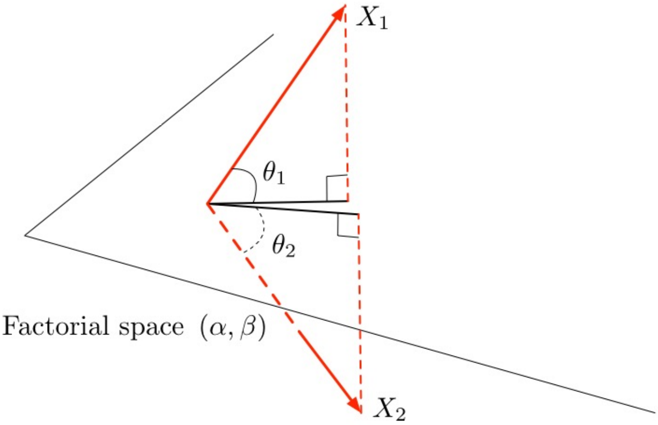
- Percentage of total inertia resumed
  Some classic values: 80%, 90%.

Quality of representation I

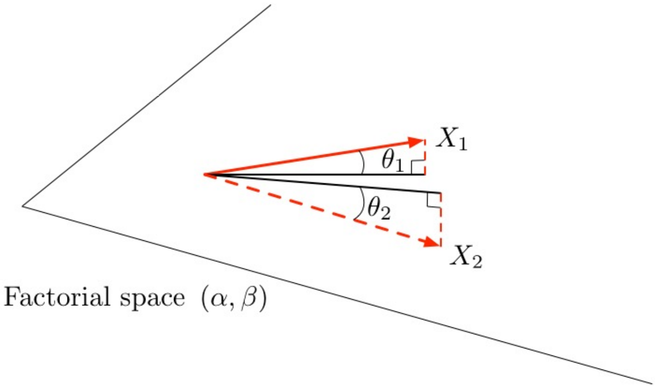
2 individuals that have close projections aren't necessarily close.

It's possible to appreciate the projection deformation by calculating the cosine of the angle between the individual and the factorial space.

Quality of representation II



Quality of representation III



Quality of representation IV

The quality of representation of X_i onto the α-th factor is:

$$\mathrm{CO2}_{\alpha}\left(i\right) = \cos^2\left(\theta_i\right) = \frac{\left(C_i^{\alpha}\right)^2}{\sum_{j=1}^{p}\left(C_i^{j}\right)^2} .$$

$\left(u_\alpha\right)_{\alpha \in \{1,\dots,p\}}$ being orthogonal, the quality of representation on a factorial space is additive:

$$\mathrm{CO2}_{\alpha+\beta}\left(i\right) = \mathrm{CO2}_{\alpha}\left(i\right) + \mathrm{CO2}_{\beta}(i) .$$

We have $\sum_{\alpha=1}^{p} \mathrm{CO2}_{\alpha}(i) = 1$.

Contribution

The contribution of the i-th individual onto the α-th factor is:

$$\mathrm{CTR}_\alpha(i) = \omega_i \frac{(C_i^\alpha)^2}{\lambda_\alpha}$$

where C_i^α is the coordinate of the i-th individual onto the α-th factor.

We have:

$$\sum_{i=1}^{n} \mathrm{CTR}_\alpha(i) = 1 \ .$$

Note that the interpretation of the contributions depends on the number of individuals.

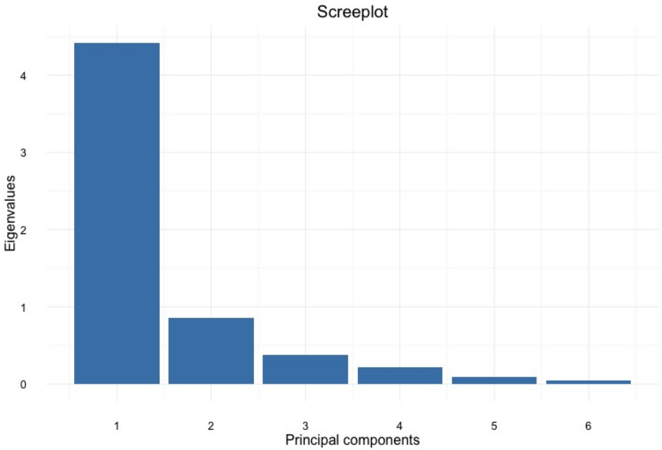
PCA "big data compatible"

- Singular Value Decomposition (SVD) is commonly used for PCA but it's also possible to use the Nonlinear Iterative PArtial Least Squares (NIPALS) algorithm which:
  - gives more numerically accurate results,
  - but is slower to calculate,
  - and suffers from a loss of orthogonality in the case of very-high-dimensional datasets with a large degree of column collinearity.
- In the case of too many individuals: covariance matrix can be incrementally computed.
- In the case of too many variables: it's possible to use methods like *very sparse random projections*.

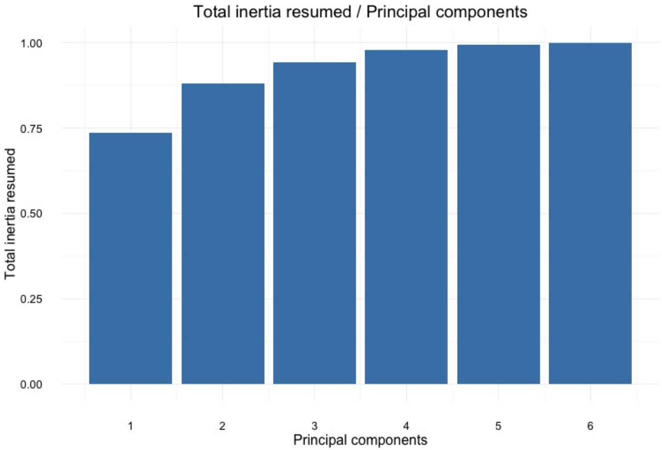
Note that there are some specific libraries in R, for exemple *bigpca*.

Table of contents

Screeplot


Screeplot
Eigenvalues
0
1
2
3
4
1
2
3
4
5
6
Principal components

Total inertia resumed



Number of principal components

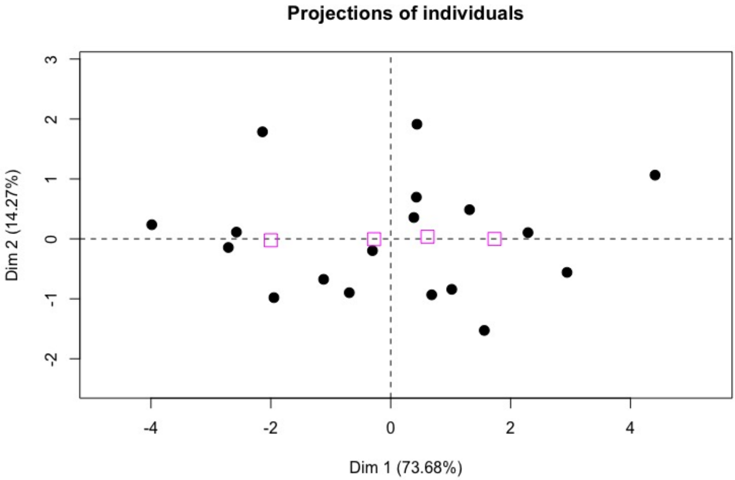
With $I = 6$:

- $l_1 = 4.42$, $\tau_1 = 73.7\%$,
- $l_2 = 0.86$, $\tau_2 = 14.3\%$.

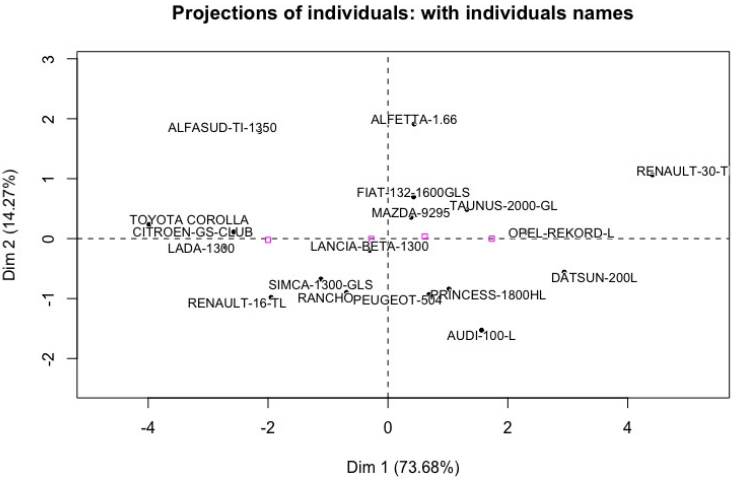
On the first factorial space:

$$\tau_{1 \oplus 2} = \frac{I_{u_1 \oplus u_2}}{I} = 87.9\% \ .$$

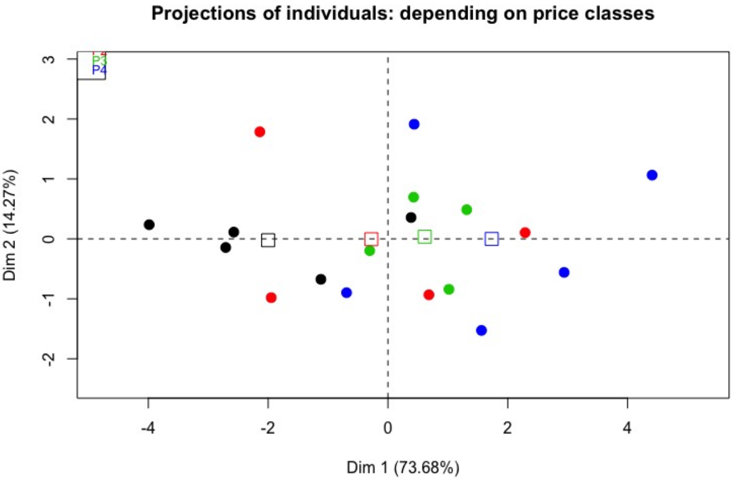
Projections on the first factorial space I



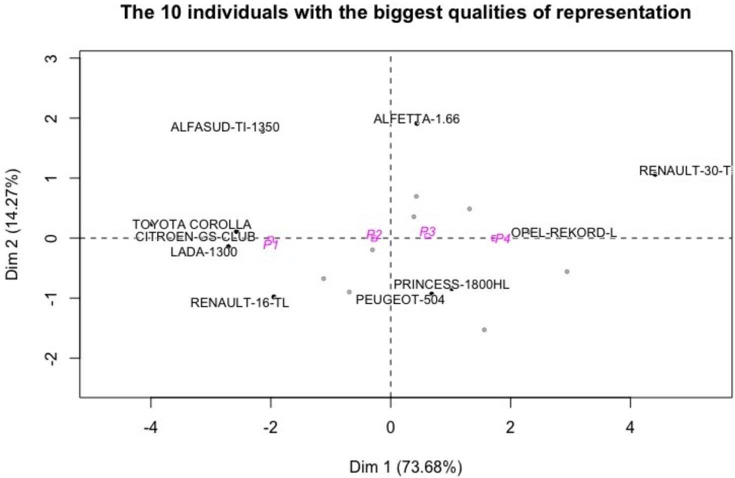
Projections on the first factorial space II



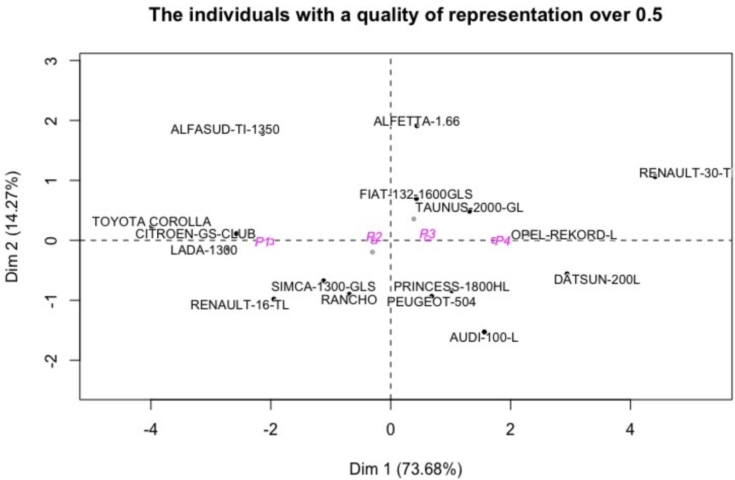
Projections on the first factorial space III



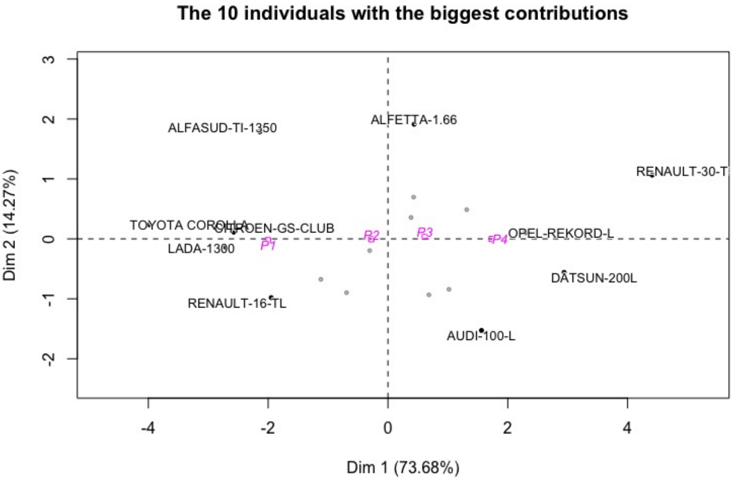
Quality of representation on the first factorial space I



Quality of representation on the first factorial space II



Contributions on the first factorial space



Case study
Problem
Method
PCA on the case study
References

References

Fénelon, J.-P. (2000). *Qu'est-ce que l'analyse des données ?* SEISAM.

Saporta, G. (2011). *Probabilités, analyse des données et statistique*. Technip, 3 edition.

Tufféry, S. (2011). *Data mining and statistics for decision making*. Wiley series in Computational Statistics. Wiley.