Introduction

Objective

The *factorial methods* aim to

- visualize, and more generally,
- handle multidimensional data.

Redundancy

Simultaneously considering many variables is a difficult problem;

Fortunately, the information provided by these variables is often redundant.

A solution

Replace the initial variables with a reduced number of new variables without losing too much information.

Principles

For example, when the variables are all quantitative, principal component analysis (PCA) seeks to solve this problem by

- considering the new variables as linear combinations of the initial variables
- uncorrelated.

Original table to synthetic table

We move from an original table X to a synthetic table with the same number of rows but a reduced number of columns C.

History

This method was first developed by K. Pearson (1900) for two variables, and then by H. Hotelling (1933), who extended it to any number of variables.

Maximizing the variance of projected data

The cloud of individuals

The data table $\boldsymbol{X}$ is an $\boldsymbol{n} \times \boldsymbol{p}$ real matrix:

- each row $\boldsymbol{x}_i^T = \boldsymbol{X}_{i\bullet}$ describes an individual with $\boldsymbol{p}$ variables
- each column $\boldsymbol{X}^j$ describes a variable with $\boldsymbol{n}$ individuals

Centering the matrix $\boldsymbol{X}$

The cloud of individuals is centered around the center of gravity of the cloud (or vector of empirical means):

$$\bar{X} = \frac{1}{n}\sum_i \boldsymbol{x}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \dots \\ X_{ip} \end{pmatrix}$$

Without loss of generality, we assume that this mean vector is the zero vector (it is sufficient to center the original matrix $\boldsymbol{X}$).

Empirical Variance The empirical variance of the cloud is the sum of the variances of each variable:

$$\hat{\sigma}^2 = \sum_{j=1}^{p} \hat{\sigma}_j^2$$

where $\hat{\sigma}j^2 = \frac{1}{n}\sum_i Xij^2$

Relationship between Empirical Variance and Inertia

$$n\hat{\sigma}^2 = \sum_i \sum_{j=1}^{p} X_{ij}^2$$

can also be interpreted as the sum of the distances from the individuals to the center of the cloud.

Projection of Individuals

Projection onto the $\boldsymbol{u}_1$ axis

The vector projection of vector $\boldsymbol{x}_i$ onto the vector line with direction vector $\boldsymbol{u}_1$ is defined as

$$c_{i1}\boldsymbol{u}_1$$

where $c_{i1} =< \boldsymbol{x}_i, \boldsymbol{u}_1 >$ is the coordinate of $\boldsymbol{x}_i$ in the $\boldsymbol{u}_1$ basis.

Projection onto the vector subspace with basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$

The vector projection of vector $\boldsymbol{x}_i$ onto the vector subspace with basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ is defined as the vector

$$c_{i1}\boldsymbol{u}_1 + \ldots + c_{id}\boldsymbol{u}_d$$

where $c_{ik} =< \boldsymbol{x}_i, \boldsymbol{u}_k >$ is the k-th coordinate of $\boldsymbol{x}_i$ in the basis.

Empirical Covariance Matrix and Diagonalization

The matrix $S = \frac{1}{n} X^t X$ is an estimation of the covariance matrix. It is a symmetric positive definite matrix. Indeed, $S = S^t$ and

$$\begin{aligned} \boldsymbol{y}^t S \boldsymbol{y} = & \quad \boldsymbol{y}^t \frac{1}{n} \sum_i X_i X_i^t \boldsymbol{y} \\ = & \quad \frac{1}{n} \sum_i (\boldsymbol{y}^t X_i)(X_i^t \boldsymbol{y}) \\ = & \quad \sum_i |\boldsymbol{y}^t X_i|^2 \geq 0 \end{aligned}$$

Interpretation

The diagonal terms are the empirical variances:

$$\hat{\sigma} j^2 = \frac{1}{n} \sum_i X i j^2$$

The off-diagonal terms are the empirical covariances:

$$\hat{\rho} j k^2 = \frac{1}{n} \sum_i X i j X_{ik}$$

Variance Projected onto an Axis

We seek to find $\boldsymbol{v}$ such that the projection of the individuals in $\boldsymbol{X}$ onto the vector $\boldsymbol{v}$ (vector projection) is maximized:

$$\{\max_{\boldsymbol{v}} \boldsymbol{v}^t S \boldsymbol{v},\ \boldsymbol{v}^t \boldsymbol{v} = 1.$$

where $S = \frac{1}{n} X^t X$.

If we express $\boldsymbol{v}$ in the (orthonormal) basis of the eigenvectors of $\boldsymbol{S}$,

$$\boldsymbol{v} = \sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j$$

then the previous problem becomes

$$\{\max_{\alpha_1,\ldots,\alpha_d} (\textstyle\sum_{j=1}^{p} \alpha_j \boldsymbol{uj})^t UDU^t (\sum j = 1^p \alpha_j \boldsymbol{u}_j),\ \sum_j \alpha_j^2$$ =

$$\{\max_{\alpha_1,\ldots,\alpha_d} (\textstyle\sum_{j=1}^{p} \alpha_j^2 \lambda_j),\ \sum_j \alpha_j^2 = 1.$$

where $\boldsymbol{\lambda}_j$ is the $\boldsymbol{j}$-th eigenvalue.

Solution

The equation gives a barycenter on the half-line of positive real numbers between λ_1 and λ_p. The maximum value of the barycenter is λ_1, and it is obtained for $\alpha_1 = 1$ and $\alpha_j = 0, \forall j \neq 1$ (because all λ_j are positive). Therefore, the solution vector is the eigenvector of S associated with the largest eigenvalue λ_1. The projection of X_i onto $\boldsymbol{u}_1$ is the first principal component:

$$C^1 = (c_{11}, \ldots, c_{n1})^t$$

$$C = XU$$

Reconstruction Formula

$$X = \sum_j C^j \boldsymbol{u}_j^t$$

The last relationship shows that the initial dataset can be reconstructed using the principal components and the principal axes. This relationship is called the reconstruction formula. If we limit ourselves to the first k $(k < p)$ terms, we obtain an approximation of the initial dataset.

Quality of Representation

Overall Quality

The overall quality of representation of the initial set X on the first k principal components is measured as the percentage of explained variance:

$$\frac{\lambda_1 + \ldots + \lambda_k}{\mathbf{trace}(S)} \times 100.$$

Relative Contribution of an Axis to an Individual

Given that the total inertia of the dataset is $\frac{1}{n}\sum_{i=1}^{p} |\boldsymbol{x}_i|^2$, the quantity $\frac{1}{n}|\boldsymbol{x}_i|^2$ represents the portion of inertia contributed by each $\boldsymbol{x}_i$.

After projection onto the axis $\boldsymbol{u}_k$, the remaining inertia is $\frac{1}{n}(C_{ik})^2$. Each of the terms $\frac{1}{n}(C_{ik})^2$ represents the portion of the initial inertia $\frac{1}{n}|\boldsymbol{x}_i|^2$ contributed by individual $\boldsymbol{i}$ and retained by the $\boldsymbol{k}$-th axis:

$$COR(i,k) = \frac{C_{ik}^2}{\|\boldsymbol{x}_i\|^2}.$$

This quantity also represents the square of the cosine of the angle formed by the individual $\boldsymbol{x}_i$ and the vector $\boldsymbol{u}_k$.

Quality of representation in a subspace

$$QLT(i,k) = \frac{\sum_{k=1}^{k} C_{ik}^2}{\|X_i\|^2} = \sum_{i=1}^{k} COR(i,\alpha).$$

Relative Contribution of an Individual to an Axis

Starting from the relationship $\lambda_\alpha = \frac{1}{n}\sum_{i=1}^{n}(C_{ik})^2$, we can decompose λ_k, the inertia preserved by the axis $\boldsymbol{u}_k$, in terms of individuals.

The portion of inertia accounted for (or explained) by individual i for the k-th axis. We have:

$$CTR(i,k) = \frac{1}{n}\frac{C_{ik}^2}{\lambda_k}.$$

Interpretation of the New Variables

Correlation Circle

Each original variable has a correlation with the new variables. These correlations are used to interpret the new variables in terms of the original ones.

$$\mathbf{cor}(X^j, C^k) = \mathbf{cor}(X^j, \boldsymbol{X}\boldsymbol{u}_k) = \frac{\mathbf{cov}(X^j, \boldsymbol{X}\boldsymbol{u}_k)}{\sqrt{\mathbb{V}(X^j)\mathbb{V}(\boldsymbol{X}\boldsymbol{u}_k)}} \& = \frac{}{\sqrt{}}$$

because

$\mathbb{V}(\boldsymbol{X}\boldsymbol{u}_k) = \boldsymbol{u}_k^\top \boldsymbol{S}\boldsymbol{u}_k = \boldsymbol{\lambda}_k$ (see previous calculation)
$\mathbf{cov}(X^j, \boldsymbol{X}\boldsymbol{u}_k) = \frac{1}{n}(X^j)^\top \boldsymbol{X}\boldsymbol{u}_k$ is the j-th coordinate of $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u}_k = \boldsymbol{\lambda}_k \boldsymbol{u}_k$. If the variables have been previously normalized, we obtain

$$\mathbf{cor}(X^j, C^k) = \sqrt{\lambda_k} u_k^j.$$

An example of PCA

The data

This is the table of grades described. Recall that these data group the grades obtained by nine students in the subjects of mathematics, science, French, Latin, and drawing:

Grades of 9 students

	math	scie	fran	lati	d.m
jean	6.0	6.0	5.0	5.5	8
aline	8.0	8.0	8.0	8.0	9
annie	6.0	7.0	11.0	9.5	11
monique	14.5	14.5	15.5	15.0	8
didier	14.0	14.0	12.0	12.5	10

	math	scie	fran	lati	d.m
andré	11.0	10.0	5.5	7.0	13
pierre	5.5	7.0	14.0	11.5	10
brigitte	13.0	12.5	8.5	9.5	12
evelyne	9.0	9.5	12.5	12.0	18

Centering the data table

The means of the five variables are respectively 9.67, 9.83, 10.22, 10.05, and 11. The column-centered table $\boldsymbol{X}$ is obtained by subtracting the corresponding mean from each column:

Centered Table

	math	scie	fran	lati	d.m
jean	-3.67	-3.83	-5.22	-4.56	-3
aline	-1.67	-1.83	-2.22	-2.06	-2
annie	-3.67	-2.83	0.78	-0.56	0
monique	4.83	4.67	5.28	4.94	-3
didier	4.33	4.17	1.78	2.44	-1

	math	scie	fran	lati	d.m
andré	1.33	0.17	-4.72	-3.06	2
pierre	-4.17	-2.83	3.78	1.44	-1
brigitte	3.33	2.67	-1.72	-0.56	1
evelyne	-0.67	-0.33	2.28	1.94	7

Variance matrix

$$\boldsymbol{S} = \frac{1}{9}\boldsymbol{X}'\boldsymbol{X}$$

Variance Matrix

	math	scie	fran	lati	d.m
math	11.39	9.92	2.66	4.82	0.11
scie	9.92	8.94	4.12	5.48	0.06
fran	2.66	4.12	12.06	9.29	0.39
lati	4.82	5.48	9.29	7.91	0.67
d.m	0.11	0.06	0.39	0.67	8.67

Principal axes of inertia

The diagonalization of the variance matrix provides the following eigenvalues (arranged in descending order):

$$\lambda_1 = 28.2533, \lambda_2 = 12.0747, \lambda_3 = 8.6157, \lambda_4 = 0.0217, \lambda_5$$

and the normalized eigenvectors or principal axes of inertia:

$$\boldsymbol{v}_1 = \begin{pmatrix} 0.51 \\ 0.51 \\ 0.49 \\ 0.48 \\ 0.03 \end{pmatrix}, \boldsymbol{v}_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.65 \\ 0.32 \\ 0.11 \end{pmatrix}, \boldsymbol{v}_3 = \begin{pmatrix} -0.05 \\ -0.01 \\ 0.11 \\ 0.02 \\ -0.99 \end{pmatrix}, \boldsymbol{v}_4 = \begin{pmatrix} \\ - \\ - \\ \\ - \end{pmatrix}$$

Quality of representation

- The inertias of the cloud projected onto the 5 axes are equal to the sum of the eigenvalues.
- The inertia of the cloud is equal to trace(S), which is also the sum of the eigenvalues, here 48.975.
- The percentages of inertia explained by each axis are therefore 57.69, 24.65, 17.59, 0.04, and 0.02.
- The percentages of inertia explained by the principal subspaces are 57.69, 82.34, 99.94, 99.98, and 100.00.
- The initial cloud is practically in a 3-dimensional space.

Principal components $C = \boldsymbol{X}V$

Composantes principales

jean	8.70	1.70	2.55	-0.15	-0.12
aline	3.94	0.71	1.81	-0.09	0.04
annie	3.21	-3.46	0.30	0.17	0.02
monique	-9.76	-0.22	3.34	-0.17	0.10
didier	-6.37	2.17	0.96	0.07	-0.19
andré	2.97	4.65	-2.63	-0.02	0.15
pierre	1.05	-6.23	1.69	0.12	0.04
brigitte	-1.98	4.07	-1.40	0.24	0.01
evelyne	-1.77	-3.40	-6.62	-0.16	-0.06

Relative contributions of axes to individuals

Relative contributions of axes to individuals

jean	0.89	0.03	0.08	0	0
aline	0.80	0.03	0.17	0	0
annie	0.46	0.53	0.00	0	0
monique	0.89	0.00	0.11	0	0
didier	0.88	0.10	0.02	0	0
andré	0.24	0.58	0.19	0	0
pierre	0.03	0.91	0.07	0	0
brigitte	0.17	0.74	0.09	0	0

Relative contributions of individuals to axes

Relative contributions of individuals to axes

jean	0.30	0.03	0.08	0.11	0.15
aline	0.06	0.00	0.04	0.04	0.02
annie	0.04	0.11	0.00	0.15	0.00
monique	0.37	0.00	0.14	0.15	0.11
didier	0.16	0.04	0.01	0.03	0.40
andré	0.03	0.20	0.09	0.00	0.25
pierre	0.00	0.36	0.04	0.07	0.02
brigitte	0.02	0.15	0.03	0.30	0.00
evelyne	0.01	0.11	0.56	0.14	0.04

Analysis in $\boldsymbol{R}^n$

The vectors $\boldsymbol{d}^\alpha$, which are the principal components associated with the different variables, are formed by the coordinates of all the variables for the same axis $\boldsymbol{v}_\alpha$, and they satisfy the relation

$$\boldsymbol{d}^\alpha = \sqrt{\lambda_\alpha}\boldsymbol{v}_\alpha.$$

We obtain

Variables

-2.73	1.97	-0.15	-0.04	0.06
-2.69	1.29	-0.04	0.08	-0.05
-2.62	-2.26	0.32	0.06	0.04

-2.58	-1.12	0.07	-0.10	-0.05
-0.16	-0.39	-2.91	0.01	0.00

Analysis in $\boldsymbol{R}^n$

It is often preferable to represent the projection of the standardized original variables. To do this, simply divide each row of the previous table by the norm of the corresponding variable

$$\|\boldsymbol{x}^j\|^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i^j)^2.$$

The $\|\boldsymbol{x}^j\|$ actually correspond to the standard deviations of the variables. We obtain 3.37, 2.99, 3.47, 2.81, and 2.94, respectively. ::: {.cell} ::: {.cell-output-display} Table: Normalized Variables

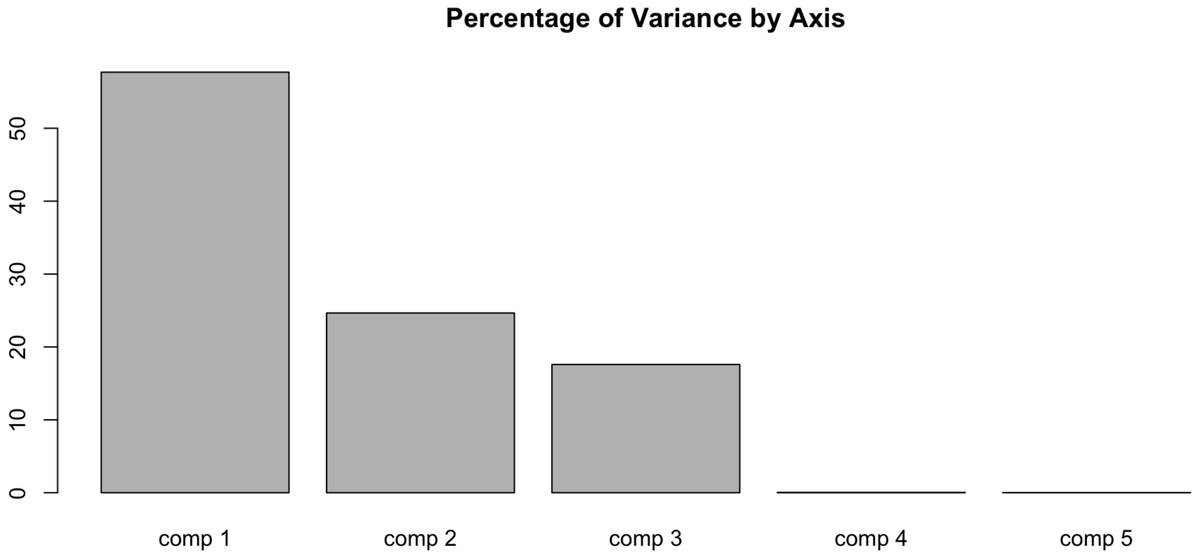
-0.81	0.58	-0.04	-0.01	0.02

-0.90	0.43	-0.01	0.03	-0.02
-0.75	-0.65	0.09	0.02	0.01
-0.92	-0.40	0.02	-0.04	-0.02
-0.06	-0.13	-0.99	0.00	0.00

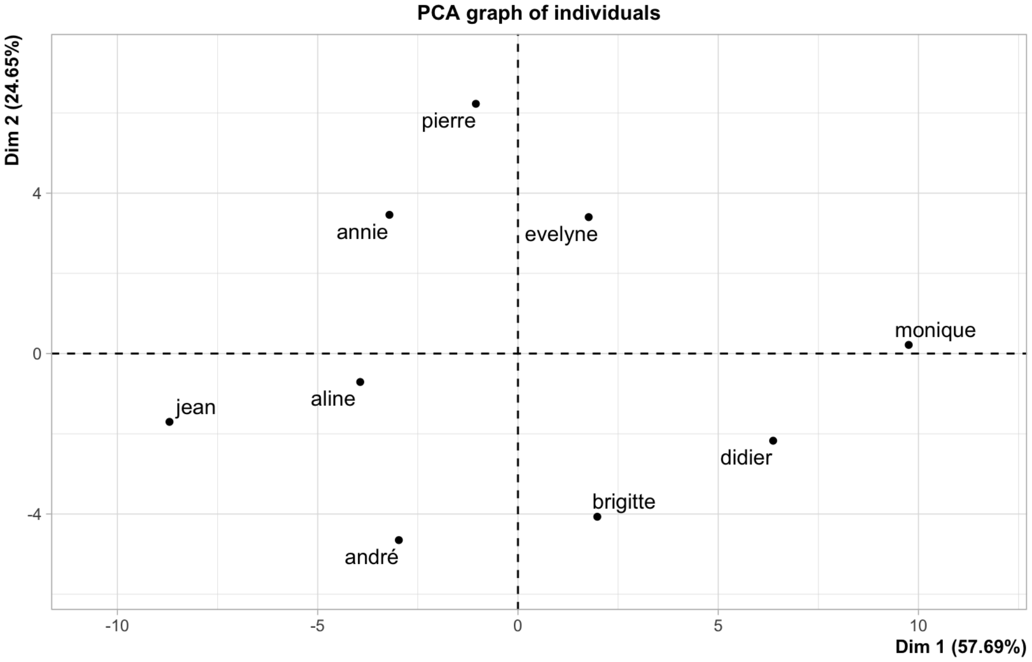
… …
… …

PCA with FactoMineR

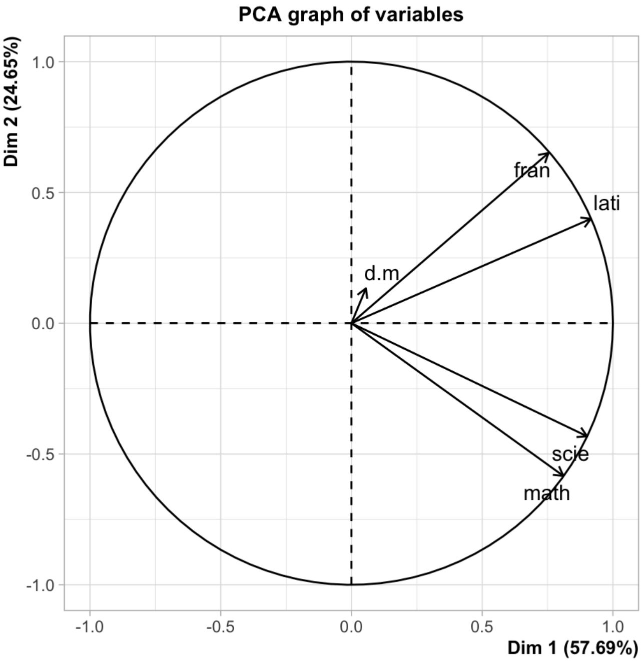
Explained Variances



Representation of Individuals



Representation of Variables

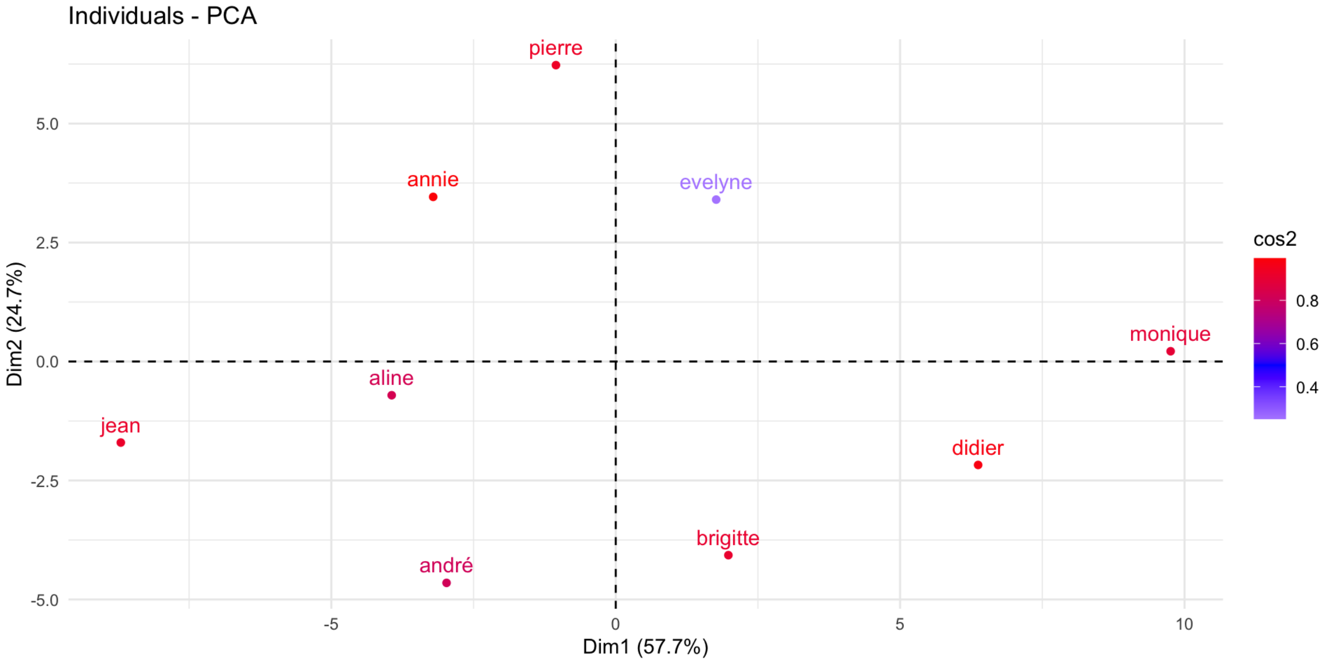


Relative contributions of axes to individuals

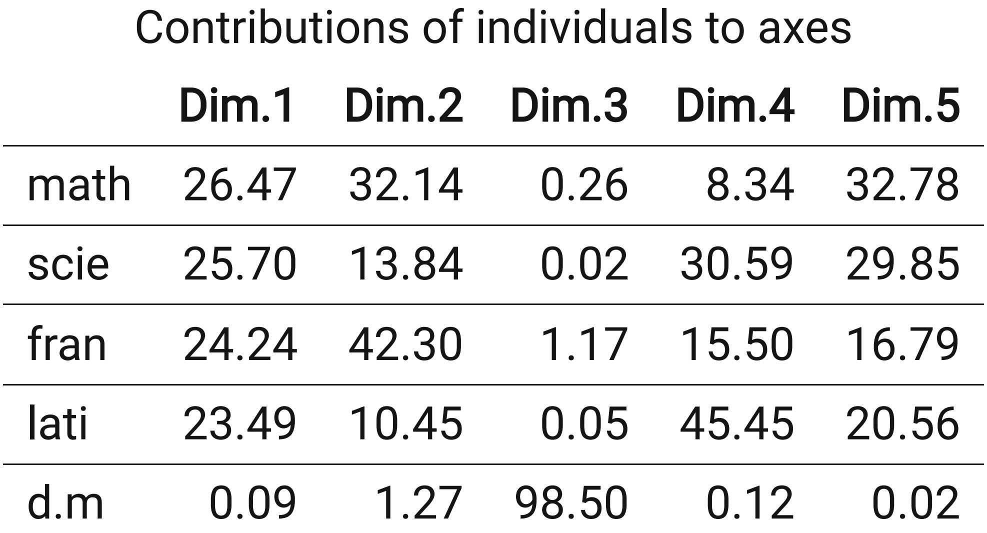
Relative contributions of axes to individuals

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
jean	0.89	0.03	0.08	0	0
aline	0.80	0.03	0.17	0	0
annie	0.46	0.53	0.00	0	0
monique	0.89	0.00	0.11	0	0
didier	0.88	0.10	0.02	0	0
andré	0.24	0.58	0.19	0	0
pierre	0.03	0.91	0.07	0	0
brigitte	0.17	0.74	0.09	0	0

Relative contributions of axes to individuals



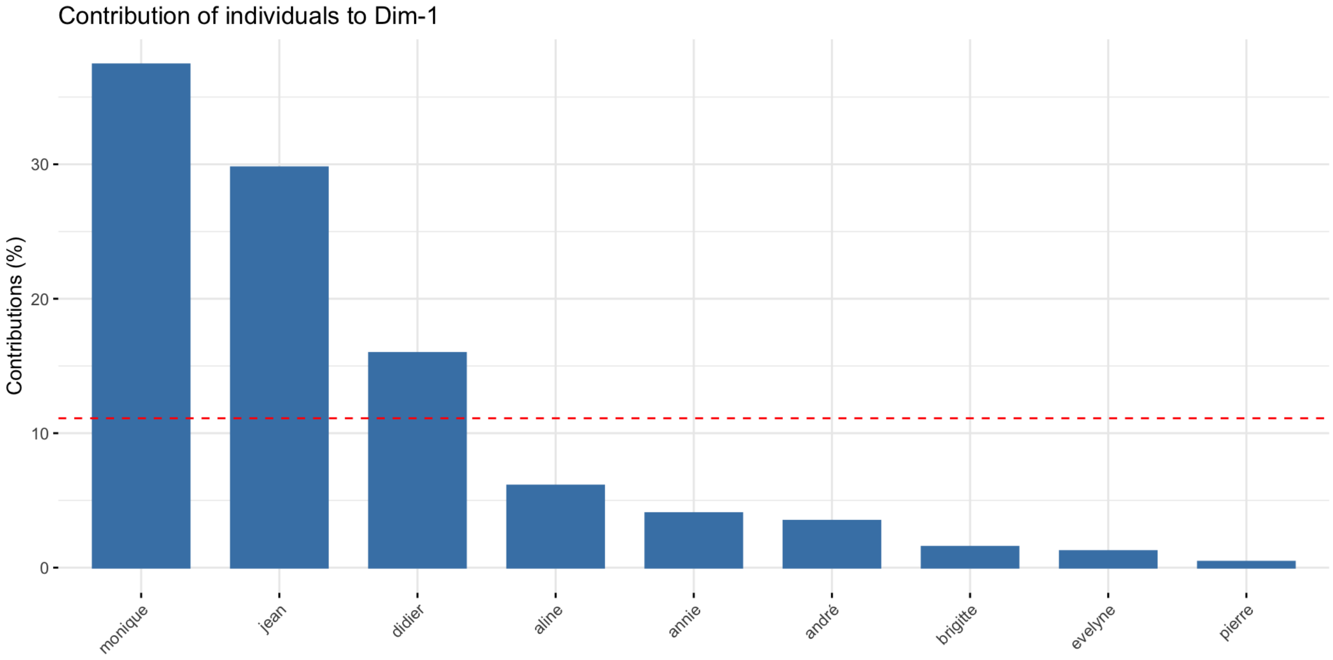
Contributions of individuals to axes



Contributions of individuals to axes

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
math	26.47	32.14	0.26	8.34	32.78
scie	25.70	13.84	0.02	30.59	29.85
fran	24.24	42.30	1.17	15.50	16.79
lati	23.49	10.45	0.05	45.45	20.56
d.m	0.09	1.27	98.50	0.12	0.02

Contributions of individuals to axes


Contribution of individuals to Dim-1
Contributions (%)
monique
jean
didier
aline
annie
andré
brigitte
evelyne
pierre