Factorial Analysis: Principal Component Analysis (PCA)

Simplified explanation

Abdelwahid Benslimane
wahid.benslimane@gmail.com

This document is a course material on principal component analysis that I have written and designed to be accessible to a wide audience. For this reason, certain aspects of the method have been deliberately omitted or simplified.

In particular, mathematical proofs/demonstrations or developments have been sidestepped in favour of a more global explanation, which cannot however be described as superficial.

The few indications in red may be ignored by those who don't have the necessary prerequisites to understand them, although reading them is relevant to others.

Let’s turn data science into a party, and let me please make you DJs (data jokeys) !

Abdelwahid Benslimane

Table of content

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Factorial analysis: context and purpose (1/1)

In real life:

- Data can be described by a very large number of variables -> analysis and interpretation challenges.
- Some of these variables may provide practically the same predictive information.
- Some variables may provide no predictive information at all (noise).

-> Negative impact on decision models.

Factorial analysis methods:

- are powerful tools for understanding the data;
- find factors (synthetic variables) that best summarize a dataset or its characteristics.

How?

- By finding the axes of maximum data dispersion

Definition

- Inertia = sum of variances of all variables

Abdelwahid Benslimane
wahid.benlimane@gmail.com

PCA (1/5)

PCA deals only with quantitative values, and its most frequent variant is the normalized PCA.

-> Variables are centred and reduced = subtracting the mean from the variables and dividing by their standard deviation (standardization). Subtracting the mean is neutral for the analysis.

Normalization process:

- D: data matrix with n rows and m columns, each row representing an individual and each column a variable
- $D_{\mathrm{i,j}}$: value of the variable j for the individual i

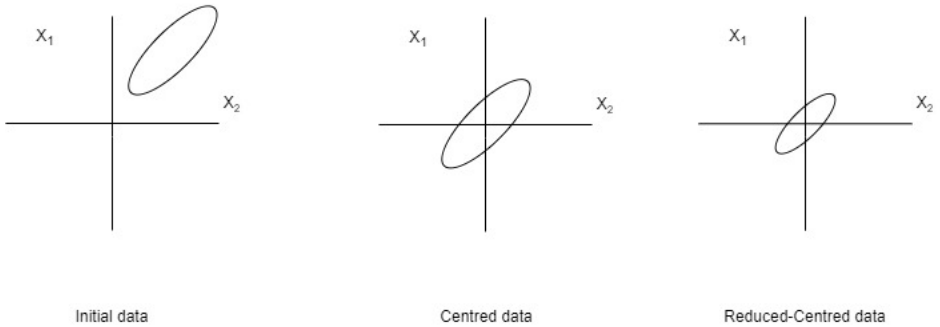
$$\frac{D_{\mathrm{i,j}} - \text{ mean } of \text{ variable j}}{\text{std. dev.} of \text{ variable j}}$$

$i = 1, 2, \ldots, n$
$j = 1, 2, \ldots, m$

- Purpose of normalization: giving the same importance to each variable regardless of the units considered

Abdelwahid Benslimane
wahid.benlimane@gmail.com

PCA (2/5)



Initial data | Centred data | Reduced-Centred data

- Example

Before normalization:

	Height (cm)	Weight (kg)	Age (years)	Annual income (€)
Individual 1	75	180	41	100 000
Individual 2	59	170	26	40 000
Indivisual 3	85	175	32	70 000

After normalization:

	Height	Weight	Age	Annual income
Individual 1	0.187	1.22	1.30	1.22
Individual 2	-1.31	-1.22	-1.14	-1.22
Indivisual 3	1.12	0	0.16	0

If all variables are measured in the same unit, it may be preferable to keep their respective variances (no standardization).

Abdelwahid Benslimane
wahid.benlimane@gmail.com

PCA (3/5)

Goal:

- To find the k axes of maximum data dispersion (principal components)
- k < < number of variables (dimensionality reduction)

-> Projecting the data onto these axes must retain as much inertia as possible.

In the example below with bidimensional data, the data are mainly dispersed along the axis associated with the first principal component (PC_1)

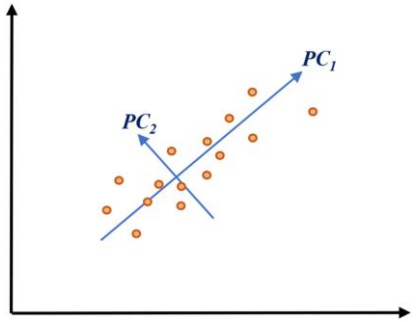


Image from https://www.researchgate.net/publication/344399773_Application_of_artificial_intelligence_models_and_optimization_algorithms_in_plant_cell_and_tissue_culture

Process:

- Calculation of the empirical correlation matrix (or covariance matrix in the case of non-normalized PCA)

D: matrix of centred data (for all individuals, we only subtract the mean for each variable, the mean therefore becomes zero for all variables)

X: normalized data matrix

Correlation matrix $R = \frac{1}{n} X^T X$ (or covariance matrix $S = \frac{1}{n} D^T D$)

n = number of data items in the data set

Abdelwahid Benslimane
wahid.benlimane@gmail.com

PCA (4/5)

- Calculation of the k (unit norm) eigenvectors of the correlation (or covariance) matrix associated with the k largest eigenvalues

Indeed, the 1st principal axis is the direction corresponding to the eigenvector u_1 associated with the largest eigenvalue λ_1.

The 2nd principal axis: direction corresponding to the eigenvector u_2 associated with the second largest eigenvalue λ_2.

And so on and so forth…

Eigenvalues correspond to the variance of data projections on the associated axes. Non-zero eigenvalues are often expressed as % of inertia (we divide each eigenvalue by the sum of the eigenvalues and then multiply by 100).

Definitions

- An eigenvector of the matrix A is any non-zero vector u verifying : $Au = \lambda u$. Eigenvectors are therefore vectors whose direction is unchanged by multiplication by the matrix under consideration. The value λ is a scalar called the eigenvalue associated with the vector.
- A unit norm vector is a vector whose norm (length) is equal to 1.

The correlation (or covariance) matrix is symmetrical by construction, and it can be shown that it is also positive (semi-)definite -> it is diagonalizable, which guarantees the existence of an eigenvector basis. This also implies that all its eigenvalues are real and positive (with the exception of those that are zero if the matrix is semi-definite) and that the eigenvectors are orthogonal, so the principal axes are orthogonal.

Abdelwahid Benslimane
wahid.benlimane@gmail.com

PCA (5/5)

How to choose the right number of principal components ?

It depends on the objective!

- For a descriptive analysis (with visualization):

We use the breaking point of the curve of eigenvalues (or % of inertia) to choose the right number of principal components (elbow method).

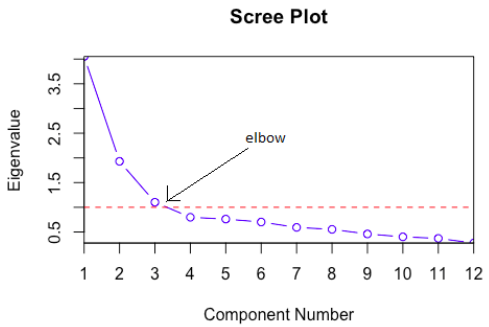


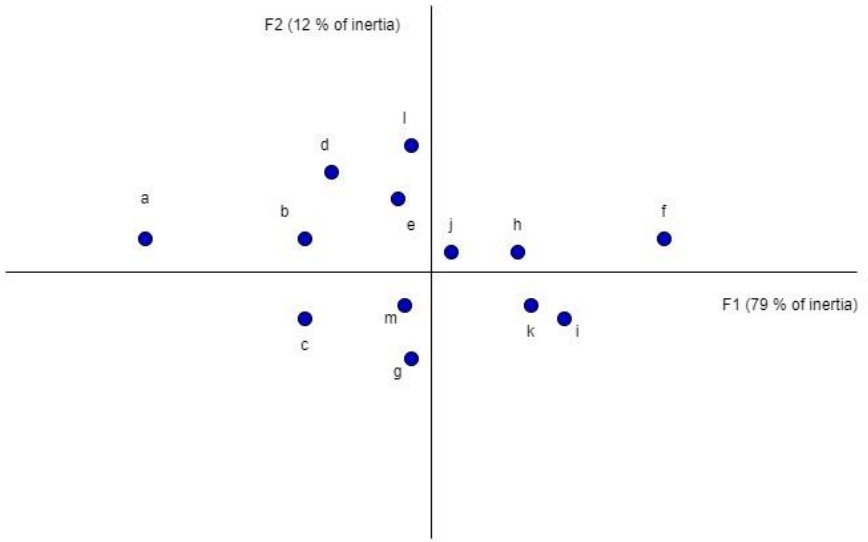
Image from https://sanchitamangale12.medium.com/scree-plot-733ed72c8608

- If PCA is used as a pre-processing step prior to the application of decision methods, the number of axes can be used as an additional parameter of the decision method.

Other approaches are possible.

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Projection of data onto factorial axes of and interpretation (1/3)



In the fictive case above, the 1st factorial axis (F1) is the one that best separates the points (it carries 79 % of the information), so we can assume that individuals **a** and **f** are very different (on most of the variables) because they are opposed on this axis.

On the other hand, individuals **i** and **k** are similar as they are very close on the graph.

Individuals very far from the origin on the first factorial axis can be considered as outliers.

The projection on the 2 main principal factorial axes/principal components is in fact the best graphical representation we can have in the current case.

What if we wanted to know what separates/brings together the individuals?

Abdelwahid Benslimane
wahid.benlimane@gmail.com

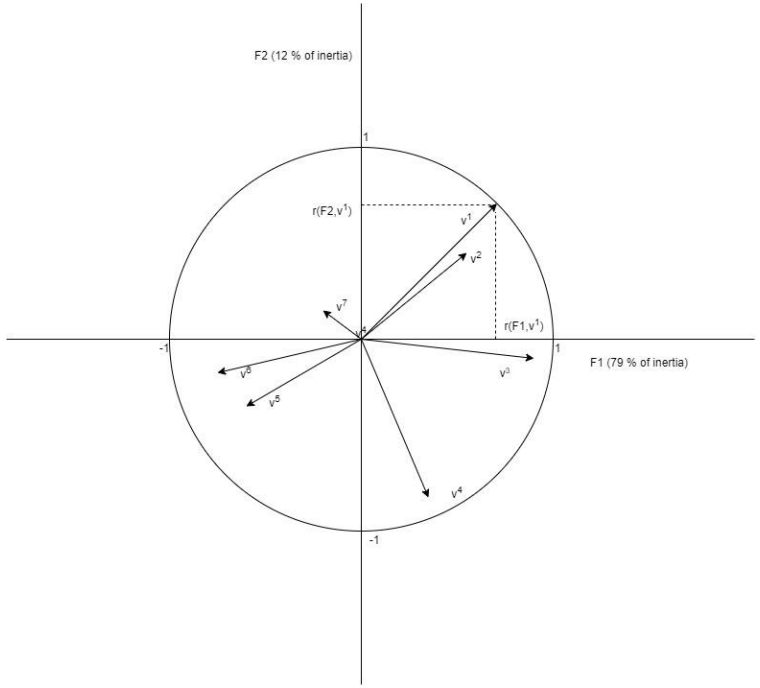
Projection of data onto factorial axes of and interpretation (2/3)

	F1	F2	…	…	Fm
Individual 1	$F1_1$	$F2_1$			Fm_1
Individual 2	$F1_2$	$F2_2$			Fm_2
…					
…					
Individual n	$F1_n$	$F2_n$			Fm_n

If we group the coordinates of all the individuals on the principal axes, we create vectors of the same dimension as the original variables. To interpret the graph of individuals, we can calculate the correlations between the initial variables (those describing the individuals in the dataset) and the main axes.

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Projection of data onto factorial axes of and interpretation (3/3)



We can then draw the correlation circle which is the graphical representation of initial variables according to their correlation coefficients with the principal components: r(Fi, v^j).

$$\frac{1}{n}\sum_{k=1}^{n} F1_k v_k^1 = C_{11} \qquad \frac{1}{n}\sum_{k=1}^{n} (F1_k)^2 = s_{F1}^2 \qquad s_{F1} = \sqrt{s_{F1}^2} \qquad \frac{1}{n}\sum_{k=1}^{n} \left(v_k^1\right)^2 = s_{v^1}^2 \qquad s_{v^1} = \sqrt{s_{v^1}^2}$$

$$r(F1, v^1) = \frac{C_{11}}{s_{F1} s_{v^1}} = \frac{< F1,\ v^1 >}{||F1||\ ||v^1||} = cos(F1, v^1)$$

If we place ourselves in the case of normalized PCA, the closer the representation of a variable is to the unit circle (= circle with radius of 1) centred at the origin, the better this variable is represented by the factorial plane (its inertia is well absorbed by the principal components). The arrows never go beyond the circle if the data is centred and reduced.

A few observations:

The variables v^1 and v^2 are highly correlated. Individuals who take strong values on the 1st factorial axis take strong values on v^3 while individuals with low values on the 1st factorial axis take high values on v^6 and vice versa.

Individuals with low values on the 2nd factorial axis take high values on v^4 and vice versa.

The variable v^7 is not well represented by the factorial plane formed by the 1st and 2nd factorial axes.

Abdelwahid Benslimane
wahid.benlimane@gmail.com