PCA- Tips

- PCA applies to numeric variables. For categorical data, other types of techniques are used, such as correspondence analysis. PCA looks for linear combinations of the variables in a way that extracts the maximum variance from the original set of variables.
- The PCA can be considered as a rotation of the coordinate system axes of the original variables to new orthogonal axes, so that these axes coincide with the direction of maximum variance of the data.
- To obtain the components, the covariance or correlation matrix is broken down into singular values. In this decomposition, the eigenvalues and the eigenvectors are determined. The principal components are linear combinations of the original variables, and there are as many as variables, with the coefficients of these combinations being the eigenvectors associated with the covariance or correlation matrix of the original variables. The ith principal component, PCi, is given by:

$$PC_i=u_{i1}X_1+\cdots+u_{ip}X_p$$

where $u_i=(u_{i1},\ldots u_{ip})$ is the eigenvector of the covariance or correlation matrix associated with the eigenvalue λ_i.

- The eigenvalues are ordered from largest to smallest and the principal component i has a variance equal to the eigenvalue λi. The first component is characterized by explaining the highest proportion of variance associated with the data, the projection of the data on it provides maximum variability. The rest of the components explain progressively smaller proportions of the variance and are characterized by not being correlated with each other. The sum of the variances of all principal components coincides with the sum of the variances of the original variables.

- Correlation between variable Xj and PCi is given by:

$$r_{ij} = \frac{u_{ij}\sqrt{\lambda_i}}{\sqrt{var(X_j)}}$$

If Xj is typified is given by:

$$r_{ij} = u_{ij}\sqrt{\lambda_i}$$

- The matrix of the rij coefficients must have the following characteristics:

Each variable must have high correlations in only one of the components. This means that it loads more on that component.

Each component must have a high correlation with a group of variables and a low correlation with the rest.

$$\lambda_i = r_{1i}^2 + \cdots + r_{pi}^2 = Var(CP_i)$$

- Contribution of variable Xj to the PCi:

$$\frac{r_{ji}^2}{r_{1i}^2 + \cdots + r_{pi}^2} = \frac{r_{ji}^2}{\lambda_i}$$

Determination of the number of components to retain

- Retain those components whose eigenvalues are greater than unity (exceed the mean of the eigenvalues), a method known as the Kaiser criterion.

- Make a sedimentation graph and decide the number of components to retain from it.

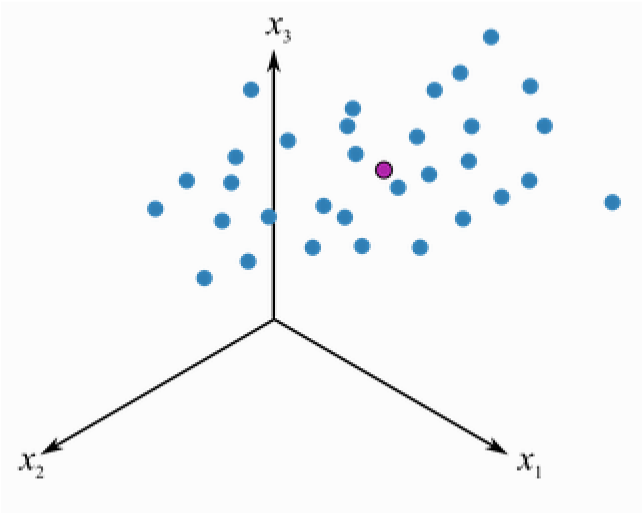
- Set a priori the minimum explained variance that you want to retain for the analysis. It is advisable that the selected components explain at least 60% of the variability in the data.

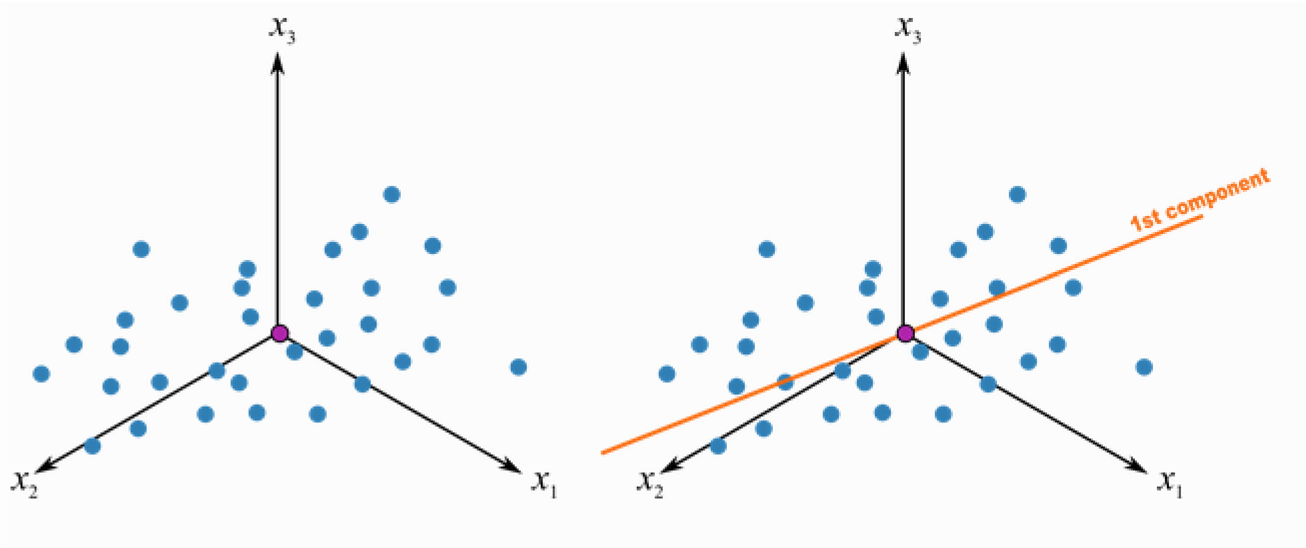
Interpretation of the components

The interpretation of the components is made from the correlation of each of them with the original variables. To do this, it is necessary to identify the variables whose correlations with the component are the highest in absolute value, which will lead to understanding its meaning and being able to assign it a name that identifies it.

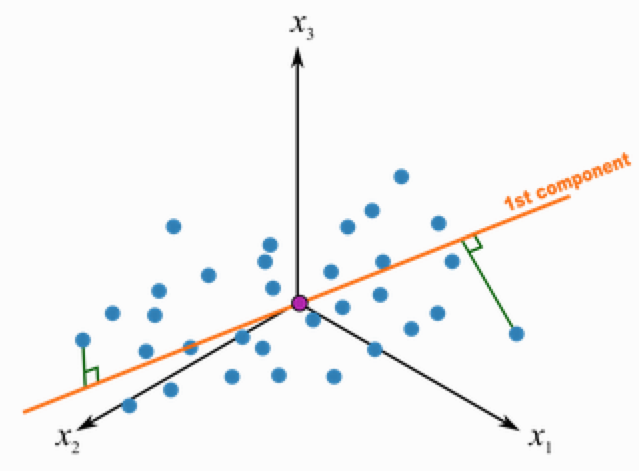
The graphical representation of the components can help their interpretation (BIPLOT). Note that the graphical representation of observations and variables is different: observations are represented by their projections, while variables are represented by their correlations. The correlation between a component and a variable estimates the information they share -> loadings, so variables can be represented as points in component space using their loadings as coordinates.

PCA Pseudo-Algorithm



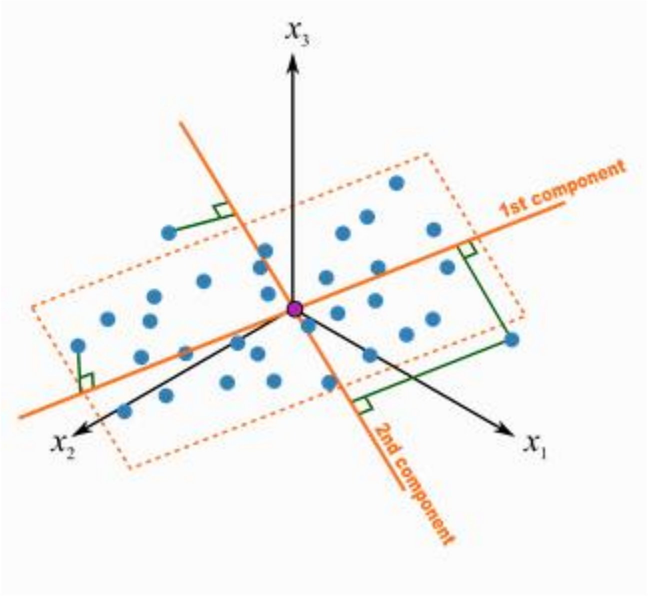


This component will be represented by a vector of loads or loadings (p×1) starting at the origin, indicating the direction of the line. On the other hand, the location of each observation on this line is obtained by projecting it at an angle of 90º:



The distance from the origin to each projection is called the score of each observation, each observation xi having its own score zi for each component m. That is, each score is a linear combination of the original value of its observation xi and the vector of loadings.

After calculating the first principal component, the second (Z2) is calculated so that it also passes through the origin and is perpendicular to the first, also maximizing the variance of the scores on this new direction. Together, Z1 and Z2 will define a plane, which is the best representation of the data in a lower dimension.



The objective is to identify the linear combinations that best represent the variables X1,...,Xp. Let (Z1,Z2,...,ZM) be M<p linear combinations of the p original variables, that is:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

where ϕ1m,ϕ2m,...,ϕpm are the loadings of the principal components (for example, ϕ11 would correspond to the first loading of the first principal component). The loadings give an idea of what weight each variable has in each component. Each vector of loadings [ϕ1m,ϕ2m,...,ϕpm] , of length equal to p, further defines the direction in space over which the variance of the data is greatest.

The linear pool is normalized so as not to inflate the variance, so:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Then, the problem is summarized as an optimization problem given by:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

This is done by calculating loadings and scores with the decomposition of the covariance (correlation) matrix in two ways: Eigen decomposition or singular value decomposition.

- Check eigenvectors and eigenvalues (video in our class materials): The eigenvectors and eigenvalues correspond to numbers and vectors associated with square matrices. Given an n×n matrix A, its eigenvector $\vec{v}$ is an n×1 matrix such that, $A\cdot\vec{v} =\lambda\cdot\vec{v}$. where the number λ is the eigenvalue, a real scalar value associated with the eigenvector.

PCA in R

→ General methods for principal component analysis

There are two general methods to perform PCA in R :

- ***Spectral decomposition***
- ***Singular value decomposition***

The function **princomp**() uses the spectral decomposition approach. The functions **prcomp**() and **PCA**()[FactoMineR] use the singular value decomposition (SVD).

According to the R help, SVD has slightly better numerical accuracy. Therefore, the function prcomp() is preferred compared to princomp().

library(stats)

- prcomp()
- princomp()

library(FactoMineR)

- PCA() -> PCA with more detailed results. Missing values are replaced by the mean of each column. Supplementary categorical variables may be included. Automatically standardizes the data.

library(factoextra)

- get_pca() -> Extracts the information about the observations and variables from a PCA analysis.
- get_pca_var() -> Extract the information about the variables.
- get_pca_ind() -> Extract the information about the observations.

Visualizations:

library(FactoMineR)

• fviz_pca_ind() -> Representation of observations on principal components.

• fviz_pca_var() -> Representation of variables on principal components.

• fviz_screeplot() -> Representation (bar graph) of eigenvalues.

• fviz_contrib() -> Represents the row/column contribution of a pca's results.

Examples:

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/#theory-behind-pca-results

https://rpubs.com/joeflorence/pca_example https://rpubs.com/aaronsc32/principal-component-analysis

PCA results for variables

Here we will show how to calculate the PCA results for variables: coordinates, cos2 and contributions:

- var.coord = loadings * the component standard deviations
- var.cos2 = var.coord^2
- var.contrib. The contribution of a variable to a given principal component is (in percentage) : (var.cos2 * 100) / (total cos2 of the component)