SMDE

Principal Component Analysis

Nihan Acar-Denizli

27 March 2020

Outline

Introduction to Principal Components Analysis (PCA)

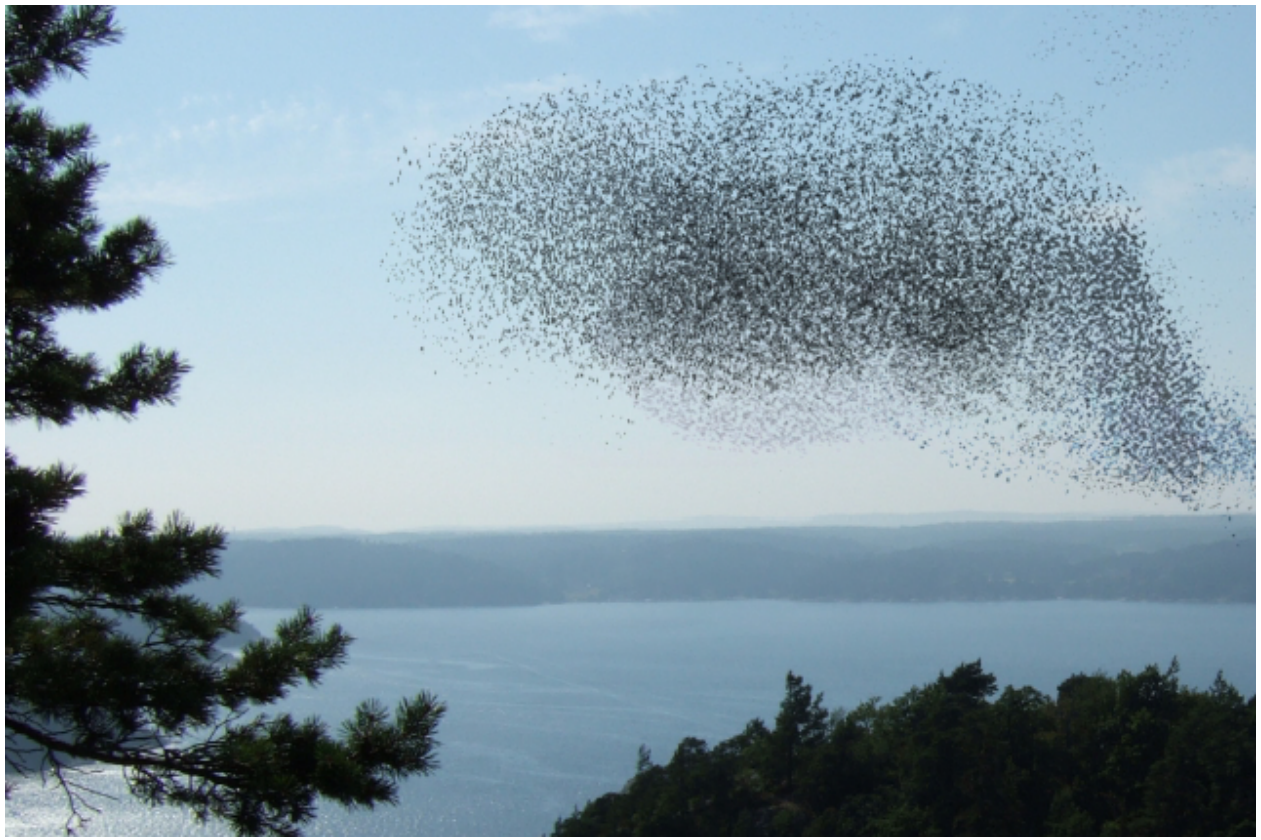
The aim of PCA is to summarize data that composes of a large set of correlated variables with a smaller number of representative variables that explain most of the variability in the original data.

PCA draw conclusions from the linear relationships between variables by detecting the principal dimensions of variability.

Principal components can be used as predictors in a regression model in place of the original correlated larger set of variables.

Introduction to Principal Components Analysis (PCA)

In the case of many variables, it is easier to describe the data using a few variables rather than all of the original variables.



The Objectives of PCA

- to analyze the correlation pattern among observed variables,
- to reduce a large number of variables to a smaller number of uncorrelated factors
- to cluster the individual observations by using the scores on factors.

Two Schools for Data Analysis

- French school of data analysis led by Jean-Paul Benzecri. Based on projections and graphical displays, representation of both rows and variables are important (Husson et al., 2011).
- English school led by based on algebra and transformation of variables. Handles the case as an optimization problem with its constraints. (Joliffe, 2002; James et al., 2017)

For detailed information see, Husson, Josse and Saporta (2016).

Why do we Need Principal Components?

Assume that we would like to visualize n observations in a p dimensional space. We might use pairs of p variables to visualize observations on the scatter plot. Then there are $\frac{p(p-1)}{2}$ possible graphs. For instance, if there are $p = 10$ variables in the data set, 45 scatterplots could be drawn.

Although, observations lives in a p-dimensional space, not all of these dimensions are equally interesting. PCA helps to represent data in a low dimensional space retaining as much as possible of the variation/information in the data set.

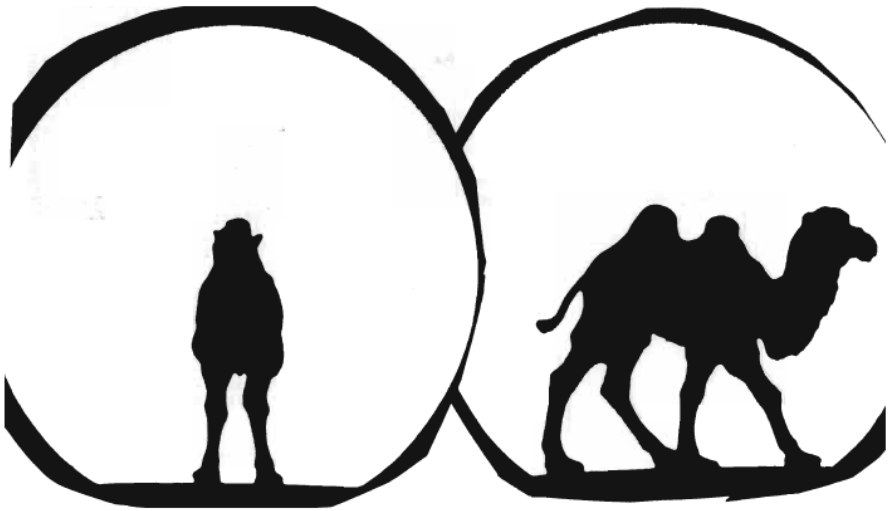


Figure: Camel or dromedary? (Illustration by J.P. Fénelon)

Steps of PCA

1. The correlation between the variables are checked.
2. The data matrix that composes of n number of observations and p number of variables is centered.
3. The covariance matrix $\mathbf{V}$ is computed.
4. The eigenvalues and eigenvectors of the covariance matrix are found.
5. m out of p number of components are chosen ($m < p$).
6. The data is projected onto the eigenvectors.
7. Uncorrelated lower dimensional of data is obtained.

What are Principal Components?

The idea behind the PCA is to describe variation in a set of correlated variables, $X_1, X_2, \ldots, X_p$, in terms of a new set of uncorrelated variables, $Z_1, Z_2, \ldots, Z_p$, each of which is a linear combination of the p number of variables.

The new variables $Z_1, Z_2, \ldots, Z_p$ are called as principal components.

There are as many principal components as the number of variables p.

The Computation of Principal Components

The first principal component of the observations, Z_1, is defined by the linear combination of

$$Z_1 = u_{11}X_1 + u_{21}X_2 + \ldots + u_{p1}X_p,$$

where $u_{11}, u_{21}, \ldots, u_{p1}$ are called and they are the elements of the loading vector $u_1 = (u_{11}, u_{21}, \ldots, u_{p1})^t$.

We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

The Computation of Principal Components

To obtain the first principal component, firstly $n \times p$ dimensional data matrix is centered by taking the difference $X_ij - \bar{X}_j$ so as to that the column means of the data matrix is zero.

If the units of measurement for variables are different, the standardized data is used in computations.
The first principal component, $z_{i1} = u_{11}X_{i1} + u_{21}X_{i2} + \ldots + u_{p1}X_{ip}$ has the largest sample variance subject to the constraint that $\sum_{j=1}^{p} u_{j1}^2 = 1$.
The principal component loading vector solves the optimization problem,

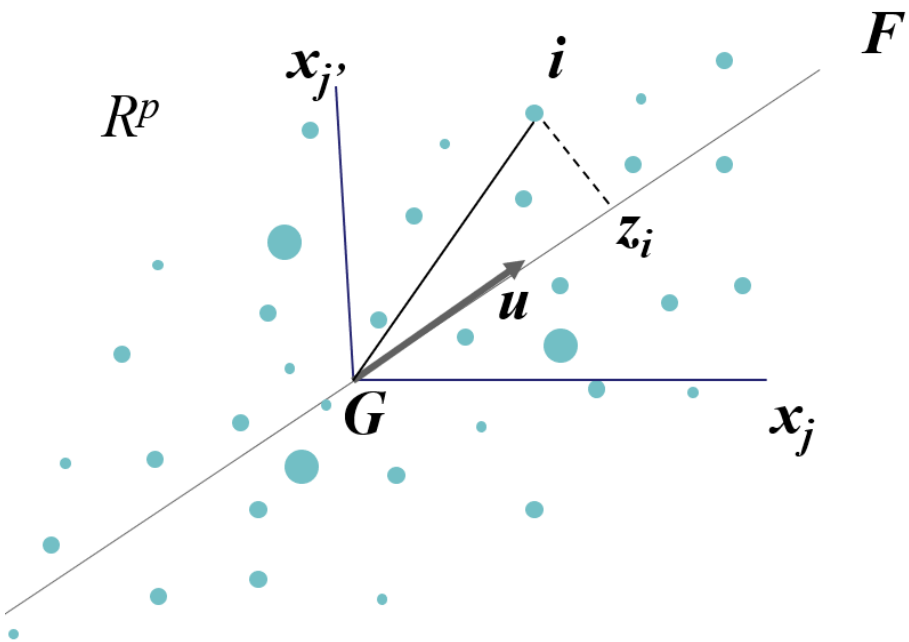
$$\text{maximize}\{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} u_{j1}x_{ij}\right)^2\} = \text{maximize}\{\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2\}$$

subject to $\sum_{j=1}^{p} u_{j1}^2 = 1$. The objective here is maximizing the sample variance of the n values of z_{i1}.

The Computation of Principal Components

The loading vector $u_1 = (u_{11}, \ldots, u_{p1})$ defines a direction in the feature space along which the data vary the most.

If we project the $X_1, \ldots, X_n$ onto this direction, the projected values are the principal component scores.



The Computation of Principal Components

The second principal component is the linear combination of $Z_2 = u_2 X_2$, which is chosen to account for as much as possible of the remaining variation, subject to two conditions that

1. it is uncorrelated with z_1 and is orthogonal to the first principal component, $u_2 u_1 = 0$.
2. it is orthonormal $u_2 u_2 = 1$.

The second principal component scores can be written in the form of,

$$z_{i2} = u_{12} x_{i1} + u_{22} x_{i2} + \ldots + u_{p2} x_{ip}$$

Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the

direction u_2 to be orthogonal (perpendicular) to the direction u_1.

Similarly, the jth component can be written in the form of the linear combination $Z_j = u_j X$, with the greatest variance subject to the conditions of orthonormality $u_j u_j = 1$, and orthogonality $u_j u_i = 0$ for $i < j$.

Three Dimensional Data to 2 Dimensions

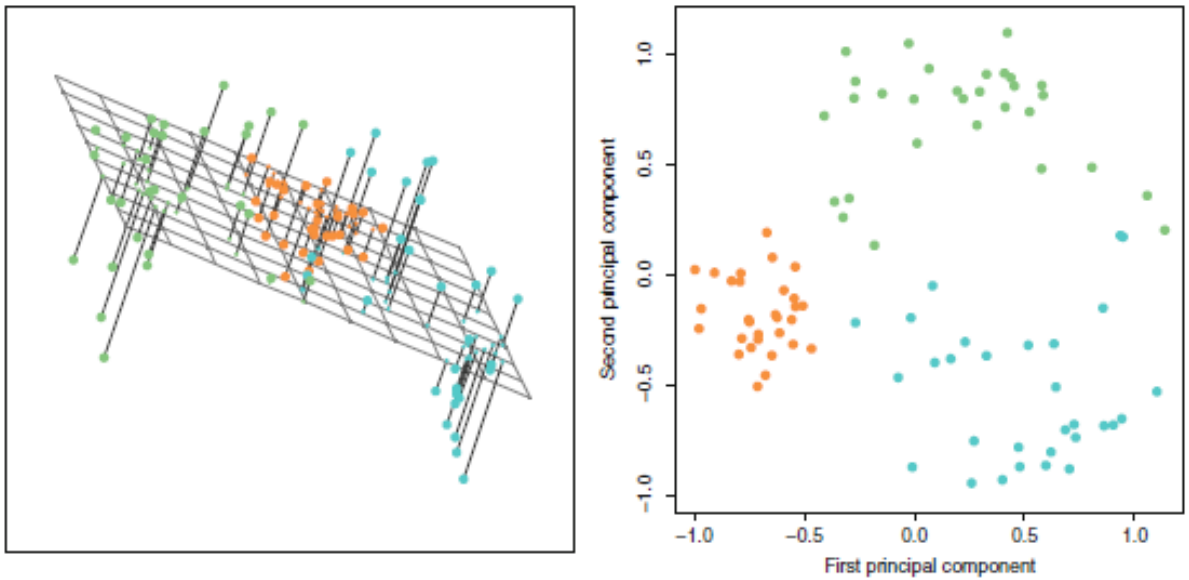


Figure: Ninety observations simulated in three dimensions. Left: the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane. Right: the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.

Loadings for USArrests Data Set

Consider the USArrests data set in HSAUR package in R. The data set consists of 50 observations and 4 variables. For 50 states the number of arrests per 100, 000 residents for each of three crimes: Assault, Murder, Rape and the percent of the population in each state living in urban areas (UrbanPop) is recorded.

The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.

The principal component loading vectors for the first two components are found as,

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Loadings for USArrests Data Set

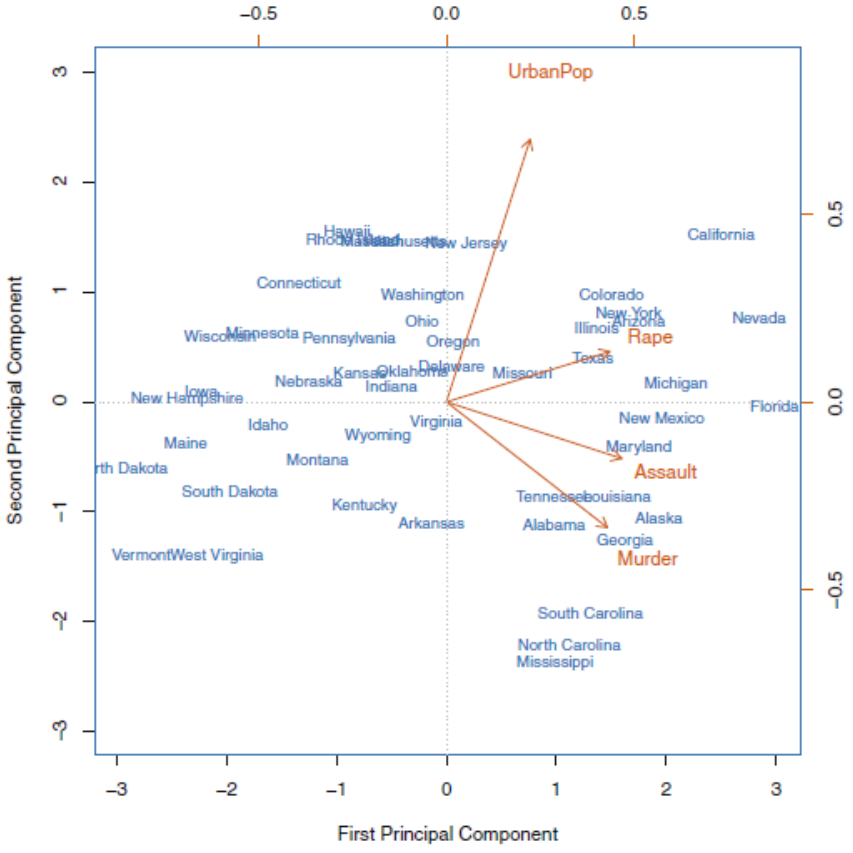


Figure: The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors.

Interpretation of Biplot

- The first loading vector places approximately equal weight on Assault, Murder, and Rape, but less weight on UrbanPop. This component corresponds to the crimes.
- The second loading vector places most of its weight on UrbanPop and much less weight on the other three features. Hence, this component roughly corresponds to the level of urbanization of the state.
- The crime-related variables (Murder, Assault, and Rape) are located close to each other, and that the UrbanPop variable is far from the other three. This indicates that the crime-related variables are correlated with each other.

Eigenvectors and eigenvalues

u_j's $(j = 1, \ldots, p)$ are called the eigenvectors of the covariance matrix of $\mathbf{V}$ associated with its jth largest eigenvalue.

Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{V}$. Then, the variance of the jth component is given by λ_j. Then the total variance of p principal components will be equal to the sum of the variances of the original variables $s_1^2, s_2^2, \ldots, s_p^2$,

$$\sum_{j=1}^{p} \lambda_j = s_1^2 + s_2^2 + \ldots + s_p^2,$$

where s_j^2 denotes the sample variance of the jth variable X_j,

$$\sum_{j=1}^{p} \lambda_j = trace(\mathbf{V}).$$

The Proportion of Explained Variation

The jth principal component accounts for a proportion P_j of the total variation of the original data,

$$P_j = \frac{\lambda_j}{trace(\mathbf{V})}.$$

The first m principal components ($m < q$) account for a proportion of,

$$p^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{trace(\mathbf{V})}.$$

These proportions are shown on a graph called "scree plot" and it is used to decide optimum number of components.

Scree Plot of USArrests

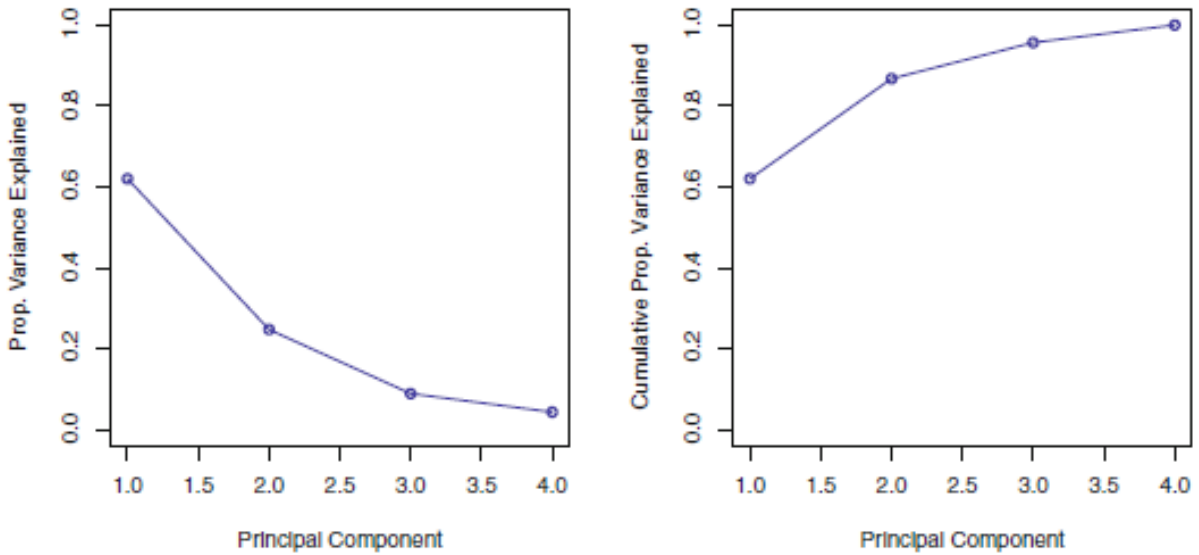


Figure: Left: A scree plot depicting the proportion of variance explained by each principal component. Right: The cumulative proportion of variance explained by the principal components.

Scaling the Variables

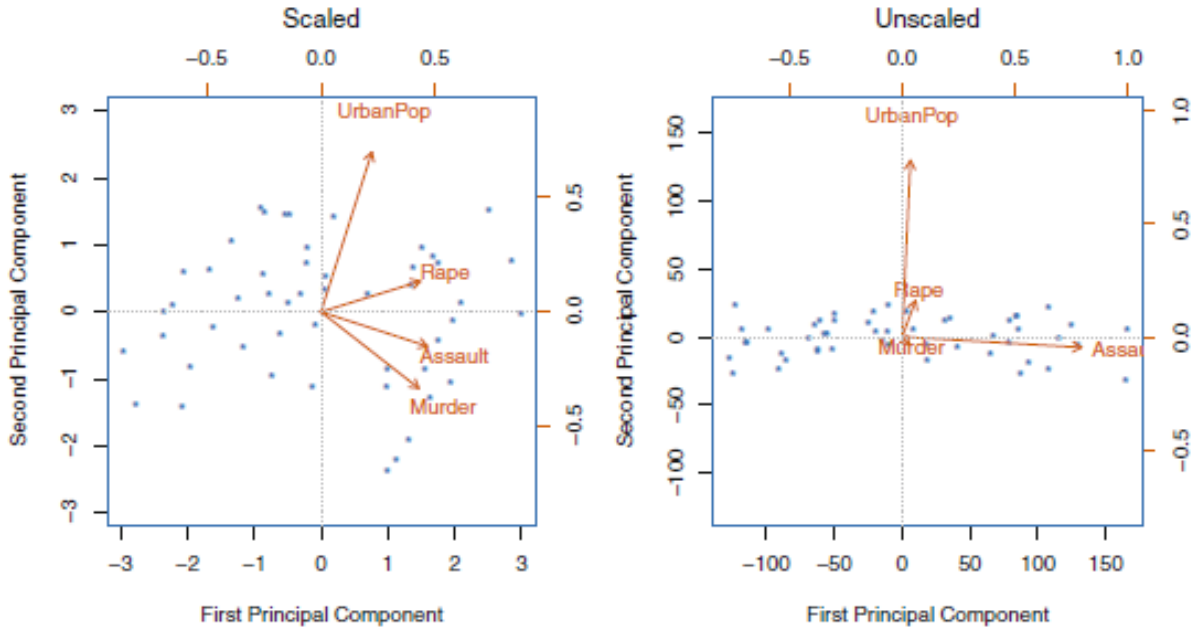


Figure: Left: The variables scaled to have unit standard deviations. Right: Principal components using unscaled data.

Scaling has a substantial effect on the results obtained.

Outline

Rotation of Components/Factors

Rotation is used to improve the interpretability and scientific utility of the solution.

Some rotation methods are:

- Orthogonal Rotation
  - Varimax: Minimize complexity of factors (most commonly used method).
  - Quartimax: Minimize complexity of variables.
  - Equamax: simplfy both variables and factors. Compromise between varimax and quartimax.
- Oblique Rotation
  - Direct oblimin
  - Direct quartimin

For detailed information, see Tabachnik and Fidell (2013).

Outline

The Assumptions

1. Normality: All variables, and all linear combinations of variables, should normally distributed.

   Normality among single variables is assessed by skewness and kurtosis, histogram or qqplots.

2. Linearity: The relationships among pairs of variables are linear.

   Linearity among pairs of variables is assessed through inspection of scatter plots.

The Assumptions

3 Absence of Outliers among Individuals: Outliers either on individual variables (univariate) or on combinations of variables (multivariate) could have more influence on the factor solution than other cases.

4 Absence of Outliers among Variables: The variables that are unrelated to other variables in the data set effect the factor results. A variable with a low squared multiple correlation with all other variables and low correlations with all important factors is an outlier among the variables.

The Assumptions

5 Factorability: A factorable data set should include several sizable correlations. It is expected that the correlation should exceed 0.30. If the correlation matrix does not include any value greater than 0.30, the use of PCA should be reconsidered.

Bartlett's (1954) test of sphericity is a notoriously sensitive test of the hypothesis that the correlations in a correlation matrix are zero.

Kaiser's (1970, 1974) measure of sampling adequacy is required to be 0.6 and above for a good PCA or Factor Analysis (FA).

REFERENCES


Everitt, B.S. and Hothorn, T.(2006). A Handbook of Statistical Analysis Using R. Chapman and Hall.


James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning with Applications in R. Springer.


Joliffe, I.T. (2002). Principal Components Analysis. Springer.


Tabachnik, B.G. and Fidell, L.S.(2013). Using Multivariate Statistics. Pearson.


Husson, F., Le, S. and Pages, J.(2011). Exploratory Multivariate Analysis by Example Using R. CRC Press.


Husson, F, Josse, J. and Saporta, G. (2016). Jan de Leeuw and the French School of Data Analysis, *Journal of Statistical Software*, 73(6).