Örebro University
Örebro University School of Business
Master's program "Applied Statistics"
Advanced level thesis II
Supervisor: Prof. Panagiotis Mantalos
Examiner: Prof. Sune Karlsson
Spring 2015

A Monte Carlo Study Comparing Three Methods for Determining the Number of Principal Components and Factors

Teodora Sheytanova
90/04/04

Abstract

A common problem in principal component analysis (PCA) and factor analysis (FA) is the choice of the number of principal components and factors.

A Monte Carlo study is performed for evaluating the accuracy of three frequently used methods for detecting the number of factors and components: Kaiser criterion (Guttman, 1954; Kaiser, 1960), acceleration factor (Cattell, 1966; Raiche, Roipel, and Blais, 2006) and parallel analysis (Horn, 1965).

The results of the analysis confirm the findings from previous papers that Kaiser criterion has the poorest performance compared with the other two analysed methods. Parallel analysis is overall the most accurate, although when the true number of factors/ components is small, acceleration factor can outperform it. The acceleration factor and Kaiser criterion perform with different accuracy for different true number of factors/ components and number of variables, whereas the parallel analysis is only affected by the sample size. Kaiser criterion tends to overestimate and acceleration factor – to underestimate the number of factors/ components. The parallel analysis shows fewer fluctuations in its accuracy and is more robust.

Considering that Kaiser criterion and the acceleration factor perform differently for different true number of factors/ components, and the parallel analysis, although generally superior, is not universal and in some cases can still be outperformed by the acceleration factor, it is recommended to combine all the methods when using PCA or FA and consider the findings from simulations performed in studies such as this one in order to draw conclusions on the true number of factors or components.

Keywords: principal component analysis; PCA; factor analysis; FA; Kaiser criterion; scree test; scree plot; acceleration factor; parallel analysis, Monte Carlo.

Contents

1. Introduction

1.1. Background

Often the analysis of statistical relationships and dependencies requires the processing of quantitative data, obtained on a large number of variables. Multivariate analysis procedures have been developed for dealing with different kinds of problems when the relationship between multiple variables is studied and large amount of information is being analyzed. Two of the most frequently used multivariate analysis procedures are the principal component analysis (PCA) and factor analysis (FA). They are used in different situations having different purpose, but can both be used for decreasing the dimensionality of the multidimensional space and are often confused with one another. PCA is used for dimensionality reduction by simply specifying a fewer number of components, which explain the majority of data variability and in effect are linear combinations of the original variables. FA on the other hand seeks to explain the variables, based on some underlying latent characteristics, which are unobservable, by sorting them into groups according to their common correlation. A common problem in PCA and FA is the choice of principal components and factors.

One of the most frequently applied method for choosing the number of factors/ components, is Kiaser criterion, also known as the Kaiser-Guttman criterion (Guttman, 1954; Kaiser, 1960), implemented in every statistical software equipped for dealing with multivariate analysis problems. Other methods are the scree test (Cattell, 1966), parallel analysis (Horn, 1965), choosing a number of components, explaining maximum variability in the data (percent variance method). Not all of them can be found in every specialized statistical software. A relevant question would therefore regard the accuracy of those methods for detecting the true number of factors and components k.

Although frequently used, many researchers don't recommend the use of Kaiser criterion. Zwick and Velicer (1986) compared 5 frequently used methods for choosing k: Kaiser criterion, scree test, Bartlett's Chi-square test (Bartlett, 1950), parallel analysis and minimum average partial procedure (MAP), used in PCA (Velicer, 1976). They concluded that the parallel analysis and the MAP procedure (for PCA) were most accurate. The scree test was also useful in combination with other methods, but they didn't recommend the use of Kaiser criterion or Bartlett Chi-square test.

The same methods in addition to the percent variance were examined by Velicer, Eaton and Fava (2000). Their recommendations were similar to the ones in the Zwick and Velicer's study.

Nevertheless, Kaiser criterion continues to be the most widely used method among the suggested (Shultz, Whitney and Zickar, 2014, p. p. 271). It is necessary for the researchers to take into consideration the extent in which Kaiser criterion erroneously estimates k as well as the orientation of the errors.

1.2. Statement of the problem and purpose of the research

All methods for choosing the number of components or factors in PCA and FA have their advantages and disadvantages and their accuracy varies in different situations. It could be difficult for researchers to choose the appropriate number of factors/ components, especially in cases where all methods point out to different solutions, because of different estimates. However, knowing these advantages or disadvantages, can make the choice of k much easier.

This thesis aims to draw attention to the problems, which may occur when using 3 of the most popular methods for choosing the number of factors/ components: Kaiser criterion, acceleration factor (Raiche, Roipel, and Blais, 2006), based on the scree test by Cattell, and parallel analysis, and compare their performance in terms of reliability and accuracy. The knowledge of the extend and direction of the errors, which may transpire, would facilitate the researcher in the choice of k. The ease of implementation shouldn't be the only criterion for choosing a method for analysis.

To conduct the analysis of the 3 methods a Monte Carlo study has been performed. The data generation process differs from the one used in previous papers. It generates a number of datasets, specifically for principal component analysis and factor analysis by controlling the true number of components and factors k. 1000 replications have been used and in each of them the value of k is estimated by implementing the three analyzed methods. Their accuracy has been assessed by summarizing the number of errors from all replications.

Different combinations for the sample size, number of variables and true number of factors and components were used for performing the analysis. The generated datasets consist of 4, 5 and 10 variables and 100, 500 and 1000 number of observations for each variable. The data was generated to have from 1 to 4 underlying factors or components, where the 4th factor was implemented in the data only when the number of variables was 10.

1.3. Organisation of the thesis

Next section of the thesis introduces the theoretical basis of PCA and FA. A short comparison of the two types of analysis is made. Section 2 also specifies the estimation process used by the three analysed methods for choosing the number of factors/ components: Kaiser criterion, acceleration factor and parallel analysis.

The methodology of the study is given in Section 3 with detailed information about the data generating process for data satisfying PCA or FA.

Section 4 is more specific in terms of the study parameters and presents the results, based on the simulations. Also, examples are given, based on real data to show how the information, obtained from the simulations can be used to facilitate the researchers in the choice of k.

Conclusions are drawn in Section 5.

2. Theoretical basis

2.1. Principal Component Analysis (PCA)

This and the following sections aim to serve as a theoretical prefix of Principal Component Analysis (PCA) and Factor analysis (FA). The theory presented here is in accordance with Richard A. Johnson and Dean W. Wichern's sixth edition "Applied Multivariate Statistical Analysis", 2007.

PCA can be used for reducing data dimension and more specifically for pinpointing k principal components – linear combinations of the original p variables: $X_1, X_2, \dots X_p$ $(k < p)$. The identified principal components can give useful insight regarding the variance-covariance structure of the data and thus PCA is particularly functional in combination with other analytical methods, such as regression or cluster analysis.

Although there are many early publications, that use basic concepts of the contemporary PCA, Pearson publication from 1901, is considered as fundamental to the advance of the method. Hotelling (1933) develops the idea further. A mathematical procedure, which uses orthogonal transformation, is defined in order to convert a set of observations on potentially dependent variables into new, linearly independent components. Only few of those components are pinpointed – the ones accounting for the greatest variability in the data. Thus, PCA defines a new set of dimensions.

A scatter plot can be created by matching the values of the observations of variables $X_1, X_2, \dots X_p$. The dimensionality of the scatter plot is p – the same as the number of variables. The first principle component is chosen so that it account for maximum amount of variability in the system. In the coordinate system it is determined by a vector, which has the direction of the greatest variability in the data. The second principle component is determined by a vector orthogonal to the first, and it accounts for the greatest variability of what's left (not in the same direction as the previous). The following principal components are determined by vectors orthogonal to the previous. The total amount of components that can be pinpointed is p - the number of orthogonal vectors. The principal components are determined by the orthogonal vectors and are uncorrelated to each other. They account for the largest possible variations in the data. The procedure allows for reduction of dimensions as only the first few components are used. The last components carry negligible amount of information about the observations and can be dropped from the analysis.

From a mathematical point of view, the vectors, which determine the principal components, are in fact the eigenvectors of the covariance matrix (or correlation matrix) of the initial observations of all variables $X_1, X_2, \dots X_p$. The eigenvector $\boldsymbol{e}$ of a matrix $\boldsymbol{A}$ is a vector that satisfies the equation:

$$\boldsymbol{A}\boldsymbol{e} = \lambda \boldsymbol{e}, \tag{1}$$

where λ is a scalar, called eigenvalue for eigenvector $\boldsymbol{e}$. There are p eigenvalues λ_i $(i = 1, \dots, p)$ for $p \times p$ - dimensional covariance/correlation matrix. For each eigenvalue, an eigenvector and respectively principal component can be defined. A normalizing restriction is put on the eigenvectors, so $\boldsymbol{e_i'e_i} = 1$ must be satisfied.

Let $\bar{\boldsymbol{x}}$ be the vector of means $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)'$ for variables $X_1, X_2, \dots X_p$, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix. Then, the eigenvalues can be calculated by solving the equation:

$$\det(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) = 0, \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix of size p. The eigenvalues are sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$), which would guarantee that the first eigenvector would have the direction of largest data variance. The eigenvectors, corresponding to each eigenvalue, are obtained according to the equation:

$$\boldsymbol{\Sigma e_i} = \lambda_i \boldsymbol{e_i}, \tag{3}$$

and the restriction $\boldsymbol{e_i'e_i} = 1$. Each eigenvector has elements e_{ik}:

$$\boldsymbol{e_1} = \begin{bmatrix} e_{11} \\ e_{21} \\ \vdots \\ e_{p1} \end{bmatrix}, \boldsymbol{e_2} = \begin{bmatrix} e_{12} \\ e_{22} \\ \vdots \\ e_{p2} \end{bmatrix}, \dots, \boldsymbol{e_p} = \begin{bmatrix} e_{1p} \\ e_{2p} \\ \vdots \\ e_{pp} \end{bmatrix}$$

The coordinates of every data point are changed according to the new dimensions, determined by the eigenvectors. The principal components are obtained by using the following linear combinations:

$$\begin{aligned} Y_1 &= e_{11}(X_1 - \bar{X}_1) + e_{21}(X_2 - \bar{X}_2) + \dots + e_{p1}(X_p - \bar{X}_p) \\ Y_2 &= e_{12}(X_1 - \bar{X}_1) + e_{22}(X_2 - \bar{X}_2) + \dots + e_{p2}\left(X_p - \bar{X}_p\right) \\ &\vdots \\ Y_p &= e_{1p}(X_1 - \bar{X}_1) + e_{2p}(X_2 - \bar{X}_2) + \dots + e_{pp}(X_p - \bar{X}_p). \end{aligned} \tag{4}$$

The following is true for the principal components:

$$Var(Y_i) = \boldsymbol{e_i'\Sigma e_i} = \lambda_i, \quad i = 1,2, \dots, p; \tag{5}$$

$$Cov(Y_i, Y_k) = \boldsymbol{e_i'\Sigma e_k} = 0, \quad i \neq k. \tag{6}$$

$$\sum_{i=1}^{p} Var(X_i) = \sigma_{11}^2 + \sigma_{22}^2 + \dots + \sigma_{pp}^2 = \sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \lambda_2 + \dots + \lambda_p \tag{7}$$

The proportion of the total variance due to k$^{\text{th}}$ component is thus: $\frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$. This proportion can be used for measuring the importance of each component.

The correlation between the principal component Y_i and variable X_k can be calculated as:

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}^2}}, \quad i, k = 1,2, \dots, p. \tag{8}$$

Principal components can be extracted from the covariance matrix of the data, but another alternative is to use the correlation matrix. Using the correlation matrix is identical to using the covariance matrix of the scaled (standardized) data. The produced eigenvectors and the interpretation of the eigenvalues differ in both cases, but using the correlation matrix has the advantage of neutralizing the effect that variables measured on large scales have on the total variance.

Only a few ($k < p$) of the p principal components are selected – the ones explaining greater part of the data variation.

2.2. Factor Analysis (FA)

Factor Analysis was developed as a method for analyzing data in psychometrics (Johnson, Richard A.; Wichern, Dean W., 2007, p. 481). This analysis is used in studies, where one assumes that the study phenomenon is affected by a small number of factors (latent variables), which cannot be measured directly, but a large number of observable variables, which are a function of those factors, exist. The idea behind Factor Analysis is to find such a function explaining variables $X_1, X_2, \dots, X_p$:

$$\boldsymbol{X}_{n\times p} = \boldsymbol{F}_{n\times k}\boldsymbol{L}_{k\times p} + \boldsymbol{E}_{n\times p}, \tag{9}$$

where $\boldsymbol{X}$ is a matrix with vectors – the observed scaled variables, and columns – the observed values for each observation. $\boldsymbol{F}$ is a matrix of factor scores with $E(\boldsymbol{F}) = \boldsymbol{0}$ and $Cov(\boldsymbol{F}) = \boldsymbol{\Phi}$. The factor scores are associated with latent variables, which are common for (explain several) observable variables. They are unobserved and must be estimated. $\boldsymbol{L}$ is a matrix with factor loadings – they can be regarded as weights of the effect of the factors on the observed variables. The higher the value of the loading, the greater the effect of the factor is on a particular variable. $\boldsymbol{E}$ is an error term matrix with $E(\boldsymbol{E}) = \boldsymbol{0}$ and $Cov(E) = \boldsymbol{\Psi} = diag(\psi_i)$, where ψ_i is the specific variance, associated with variable i. Each column in $\boldsymbol{E}$ specifies a latent variable, which affects variable i, but unlike the common factors in $\boldsymbol{F}$, those in $\boldsymbol{E}$ are unique for each variable i. Also, the errors must be uncorrelated and independent from the factor scores: $Cov(\boldsymbol{F}, \boldsymbol{E})$=**0**.

Factor analysis falls in two categories (Williams, Brown et al. 2010):

- Exploratory – when one does not know the number of factors (latent variables) or how these factors affect the study phenomenon.
- Confirmatory – when a particular hypothesis about the number of factors and their effect is examined.

In general, the steps for performing factor analysis include:

- Determining the number of uncorrelated factors ($k < p$), based on the p variables.
- Estimating the factor loading **L** and factor scores **F**. By identifying the loadings values, one can draw conclusions on which factors affect which variables and thus identify a group of variables X, affected by the same factor. The variables within each group are highly correlated among themselves, but not so much with variables from the other groups, affected by other factors.

- Interpreting the results with respect to the subject area.

The following requirements must be met for applying factor analysis:

- Random data;
- The data must be measured on interval or ordinal scales;
- The sample data consists of at least 50 observations according to Sapnas and Zeller, 2002 or better no less than 100 cases as pointed out by Hair et al., 1995.

The covariance matrix for $\boldsymbol{X}$ can be expressed as:

$$\boldsymbol{\Sigma} = E[\boldsymbol{X}'\boldsymbol{X}] = E[(\boldsymbol{FL}+\boldsymbol{E})'(\boldsymbol{FL}+\boldsymbol{E})] = \boldsymbol{L}'E[\boldsymbol{F}'\boldsymbol{F}]\boldsymbol{L} + \boldsymbol{L}'E[\boldsymbol{F}'\boldsymbol{E}] + E[\boldsymbol{E}'\boldsymbol{F}]\boldsymbol{L} + E[\boldsymbol{E}'\boldsymbol{E}] =$$

$$= \boldsymbol{L}'\boldsymbol{I}\boldsymbol{L} + \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{\Psi} = \boldsymbol{L}'\boldsymbol{L} + \boldsymbol{\Psi} \qquad (10)$$

Let

$$\boldsymbol{L} \equiv \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \ldots & l_{kp} \end{bmatrix}. \qquad (11)$$

Then,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sum_{q=1}^{k} l_{q1}^2 + \psi_1 & \sum_{q=1}^{k} l_{q1}l_{q2} & \cdots & \sum_{q=1}^{k} l_{q1}l_{qp} \\ \sum_{q=1}^{k} l_{q1}l_{q2} & \sum_{q=1}^{k} l_{q2}^2 + \psi_2 & \cdots & \sum_{q=1}^{k} l_{q2}l_{qp} \\ \vdots & & \ddots & \vdots \\ \sum_{q=1}^{k} l_{q1}l_{qp} & \sum_{q=1}^{k} l_{qp}l_{q2} & \cdots & \sum_{q=1}^{k} l_{qp}^2 + \psi_p \end{bmatrix}, \qquad (12)$$

or the total variance for variable X_i can be expressed as: $\sigma_{ii} = \sum_{q=1}^{k} l_{qi}^2 + \psi_i$, where $\sum_{q=1}^{k} l_{qi}^2 = h_i^2$ is referred to as communality and ψ_i – as specific variance. Additionally, $\sigma_{ki} = \sum_{q=1}^{k} l_{qi}l_{qk}$ is the covariance between X_i and X_k.

It can be shown that the covariance between the observable and latent variables is in fact the matrix of factor loadings:

$$Cov(\boldsymbol{X},\boldsymbol{F}) = E[(\boldsymbol{FL}+\boldsymbol{E})'\boldsymbol{F}] = \boldsymbol{L}'E[\boldsymbol{F}'\boldsymbol{F}] + E[\boldsymbol{E}'\boldsymbol{F}] = \boldsymbol{L}'\boldsymbol{I} + \boldsymbol{0} = \boldsymbol{L}', \qquad (13)$$

which leads to the following expression for the correlation between X_i and F_q, given that $Cov(\boldsymbol{F}) = \boldsymbol{\Phi} = \boldsymbol{I}$:

$$\rho(X_i,F_q) = \frac{Cov(X_i,F_q)}{\sqrt{h_i^2 + \psi_i}} = \frac{l_{iq}}{\sqrt{h_i^2 + \psi_i}}. \qquad (14)$$

For easier interpretation of the effect of the factors on the observable variables, an orthogonal rotation of the axes can be applied. The most commonly applied type of rotation is 'varimax', suggested by Kaiser (1958) and it scales the loadings, making them either small or large, but the factors keep their properties for explaining the data. The values of the large factor loading increase and those of the small – decrease. A disadvantage of the rotation is the fact the factors become more correlated.

The varimax rotation focuses on finding an orthogonal matrix $\boldsymbol{T}_{k\times k}$, based on:

$$\max_{\boldsymbol{T}} Var(\boldsymbol{TL}). \tag{15}$$

Then a new matrix of factor loadings $\boldsymbol{L}^*$ is created: $\boldsymbol{L}^* = \boldsymbol{TL}$. The covariance matrix can be expressed as $\boldsymbol{\Sigma} = \boldsymbol{L}^{*\prime}\boldsymbol{L}^* + \boldsymbol{\Psi}$, because $\boldsymbol{L}^{*\prime}\boldsymbol{L}^* + \boldsymbol{\Psi} = \boldsymbol{L}'\boldsymbol{T}'\boldsymbol{TL} + \boldsymbol{\Psi} = \boldsymbol{L}'\boldsymbol{L} + \boldsymbol{\Psi}$.

There are two main methods for the estimation of $\boldsymbol{L}$ and **F**: principal component solution and maximum likelihood estimation (MLE), which can be applied under normality – only when $\boldsymbol{F}$ and $\boldsymbol{E}$ are multinormal.

The principal component solution for the factor loadings is:

$$\boldsymbol{L} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e_1} \\ \sqrt{\lambda_2}\boldsymbol{e_2} \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{e_k} \end{bmatrix}, \tag{16}$$

Where, similarly to PCA, λ_i are the eigenvalues with the corresponding eigenvector $\boldsymbol{e_i}$ for the covariance matrix $\boldsymbol{\Sigma}$ or the correlation matrix $\boldsymbol{R}$. In reality, an adjusted correlation matrix is used, for which the diagonals are modified to account for the unique factors in matrix $\boldsymbol{E}$. Thus the factor analysis deals only with the common factors in the data. To do this however, first the communalities must be estimated. Their estimators are the squared multiple correlations (SMC) of the original correlation matrix $\boldsymbol{R}$ (Guttman, 1956). The diagonals of $\boldsymbol{R}$ are replaced with the communalities estimations: $1 - [diag(\boldsymbol{R}^{-1})]^{-1}$.

MLE maximizes the likelihood function of the covariance matrix:

$$logL(\boldsymbol{\Sigma}) = -\frac{1}{2}n.\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}n.tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}). \tag{17}$$

Factor analysis can solve the following problems:

- Classification of the studied variables and their grouping in factors on the basis of their correlation;
- Discarding of the data, which gives negligible information;
- Formulating adequate factor models, explaining in high percentage the variability in the data;
- Identifying factors, which are independent to each other and are suitable for use in statistical analysis such as regression.

2.3. Similarity and difference between PCA and FA

Principal component analysis and factor analysis are similar methods, which are often mistaken for one another as they both reduce the number of available variables. Other similarities include that they require the variables to be measure on interval or ordinal scales. The analyzed data in both cases must be random and assume linear relationship between variables.

Differences include:

1. The variables in factor analysis are a function of the factors, whereas the principal components are an outcome of the variables.
2. The purpose of factor analysis is to sort the studied variables into groups with high within-group correlation. Thus, FA accounts only for the common variance. On the other hand, principal component analysis aims to reduce the number of variables, through the components, that retain maximum amount of total variance.
3. Factor analysis discriminates between common and unique variance, and principal component analysis doesn't.
4. A common misconception is to use an unadjusted correlation matrix in factor analysis, which is used in principal component analysis. In factor analysis the diagonal of the correlation matrix, which is decomposed does not consist of ones. The values of the diagonal are replaced with estimations of the commonalities. This is a consequence from point 3. The eigenvalues of the adjusted correlation matrix can be negative.

2.4. Methods for determining the number of components and factors in PCA and FA

A number of methods exist for determining the number of principal components and factors to be retained. This thesis compares three of the most commonly used methods: Kaiser criterion (Guttman, 1954; Kaiser, 1960), parallel analysis and scree test acceleration factor. Other methods include: scree test (Cattell, 1966), choosing factors/ components explaining a particular ratio of the total variance (e.g. 80%), Bartlett Chi-squared test (Bartlett, 1950).

2.4.1. Kaiser criterion

Kiaser criterion differs in PCA and FA. Originally, it was created for principal component analysis. In PCA the Kaiser criterion drops the components, for which the eigenvalues are less than 1 (when the data is standardized). Greater than 1 eigenvalue suggests that the corresponding component explains more variance than a single variable, given that a variable accounts for a unit of variance (Beavers, 2013). This can be inferred from properties (5) and (7). Therefore, the component in question can be used for reducing the number of variables. On the contrary components with eigenvalues less than 1 would not be useful for reducing the dimensionality of the data. Therefore, the rule derived by Kaiser and Guttman (Guttman, 1954; Kaiser, 1960) would be to select those components Y_i, for which $\lambda_i > 1$.