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 ........................................................................................................................................... 3
   1.1. Background ........................................................................................................................................ 3
   1.2. Statement of the problem and purpose of the research ................................................................. 4
   1.3. Organisation of the thesis .................................................................................................................. 4
2. Theoretical basis ...................................................................................................................................... 5
   2.1. Principal Component Analysis (PCA) ............................................................................................... 5
   2.2. Factor Analysis (FA) ......................................................................................................................... 7
   2.3. Similarity and difference between PCA and FA............................................................................... 10
   2.4. Methods for determining the number of components and factors in PCA and FA ...................... 10
      2.4.1. Kaiser criterion ......................................................................................................................... 10
      2.4.2. Scree test acceleration factor .................................................................................................. 11
      2.4.3. Parallel Analysis ....................................................................................................................... 12
3. Methodology: Design of the Monte Carlo simulation ............................................................................ 12
   3.1. Defining a set of data generating processes (DGP) ........................................................................ 12
      3.1.1. DGP used for detecting the number of principal components .............................................. 12
      3.1.2. DGP used in detecting the number of factors ........................................................................ 13
   3.2. Generating r independent replications ............................................................................................ 14
   3.3. Estimations ....................................................................................................................................... 14
4. Implementation ....................................................................................................................................... 15
   4.1. Data and model of the data generating process .............................................................................. 15
   4.2. Results ............................................................................................................................................. 15
      4.2.1. Detected number of principal components ............................................................................. 15
      4.2.2. Detected number of factors .................................................................................................... 20
   4.3. Example “Places” dataset .............................................................................................................. 25
      4.3.1. Number of principal components .......................................................................................... 25
      4.3.2. Number of factors .................................................................................................................... 28
5. Summary and Conclusion ..................................................................................................................... 32
   5.1. Summary .......................................................................................................................................... 32
   5.2. Conclusion .................................................................................................................................... 32
6. Recommendations .................................................................................................................................. 33
7. Bibliography .......................................................................................................................................... 34
8. Appendix ............................................................................................................................................... 35
   8.1. Frequency plots ............................................................................................................................... 35
      8.1.1. Detected number of components in all 1000 replications ...................................................... 35
      8.1.2. Detected number of factors in all 1000 replications .............................................................. 42
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: \( X_1, X_2, \ldots, X_p \) \((k < p)\). 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 \( X_1, X_2, \ldots, X_p \). 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 \( X_1, X_2, \ldots, X_p \). The eigenvector \( e \) of a matrix \( A \) is a vector that satisfies the equation:

\[
Ae = \lambda e,
\]

where \( \lambda \) is a scalar, called eigenvalue for eigenvector \( e \). There are \( p \) eigenvalues \( \lambda_i \) \((i = 1, \ldots, p)\) 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 $e_i' e_i = 1$ must be satisfied.

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

$$\det(\Sigma - \lambda I) = 0,$$

(2)

where $I$ is the identity matrix of size $p$. The eigenvalues are sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$), 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:

$$\Sigma e_i = \lambda_i e_i,$$

(3)

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

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

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:

$$Y_1 = e_{11}(X_1 - \bar{X}_1) + e_{21}(X_2 - \bar{X}_2) + \cdots + e_{p1}(X_p - \bar{X}_p)$$
$$Y_2 = e_{12}(X_1 - \bar{X}_1) + e_{22}(X_2 - \bar{X}_2) + \cdots + e_{p2}(X_p - \bar{X}_p)$$
$$\vdots$$
$$Y_p = e_{1p}(X_1 - \bar{X}_1) + e_{2p}(X_2 - \bar{X}_2) + \cdots + e_{pp}(X_p - \bar{X}_p).$$

(4)

The following is true for the principal components:

$$\text{Var}(Y_i) = e_i' \Sigma e_i = \lambda_i, \quad i = 1, 2, ..., p;$$
$$\text{Cov}(Y_i, Y_k) = e_i' \Sigma e_k = 0, \quad i \neq k.$$

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

(7)

The proportion of the total variance due to $k^{th}$ component is thus: $\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$. 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:

$$\rho_{Y_i X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}^2}}, \quad i, k = 1, 2, ..., p.$$
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 $X_1, X_2, \ldots, X_p$:

$$X_{n\times p} = F_{n\times k}L_{k\times p} + E_{n\times p},$$

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

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 $X$ can be expressed as:

$$
\Sigma = E[X'X] = E[(FL + E)'(FL + E)] = L'E[F'F]L + L'E[E'E] + E[E'F]L + E[E'E] = L'IL + 0 + 0 + \Psi = L'L + \Psi
$$

(10)

Let

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

(11)

Then,

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

(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 $\psi_i$ 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:

$$
Cov(X, F) = E[(FL + E)'F] = L'E[F'F] + E[E'F] = L'I + 0 = L',
$$

(13)

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

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

(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 $T_{k \times k}$, based on:

$$\max_T Var(TL).$$

(15)

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

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

The principal component solution for the factor loadings is:

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

(16)

Where, similarly to PCA, $\lambda_i$ are the eigenvalues with the corresponding eigenvector $e_i$ for the covariance matrix $\Sigma$ or the correlation matrix $R$. In reality, an adjusted correlation matrix is used, for which the diagonals are modified to account for the unique factors in matrix $E$. 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 $R$ (Guttman, 1956). The diagonals of $R$ are replaced with the communalities estimations: $1 - [\text{diag}(R^{-1})]^{-1}$.

MLE maximizes the likelihood function of the covariance matrix:

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

(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 measured on interval or ordinal scales. The analyzed data in both cases must be random and assume linear relationships 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

Kaiser 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 $\lambda_i > 1$. 

2.4.2. Parallel analysis

Parallel analysis is a method to decide on the number of factors to retain. A random sample of size $n$ is created and a principal component analysis is run on this sample. The eigenvalues of this sample are compared to the eigenvalues of the original data. The number of factors retained is the number of eigenvalues in the sample that are greater than the eigenvalues of the original data.

2.4.3. Scree test acceleration factor

The scree test is a graphical method to determine the number of factors to retain. The scree plot is a graph of the eigenvalues of the correlation matrix. The acceleration factor is a measure of how much the eigenvalues change. The number of factors to retain is the point where the plot levels off.
For FA, the eigenvalue threshold is 0, since the eigenvalues, obtained from the adjusted correlation (or covariance) matrix can be negative. The reason for this is that the adjusted matrices do not have full rank – the common variance is less than the total. The negative eigenvalues “cannot be safely interpreted as partitions of the common variance” (Lorenzo-Seva, 2013). Since factor analysis focuses only on the common variance, only factors with corresponding positive eigenvalues are selected for rotation ($\lambda_i > 0$).

### 2.4.2. Scree test acceleration factor

Unlike the exact rule, determined by the Kaiser criterion, the Cattel’s scree test is more subjective and it relies on visual interpretation of the eigenvalues curve on the scree plot. ‘Scree’ is a geological term referring to fragment of rocks, gathered at the bottom of a mountain or a cliff. Plotting the eigenvalues of the factors/components to the corresponding factors gives a curve that visually resembles the outlines of a cliff or a mountain, since the first eigenvalues are largest. The more factors/components, the lower the eigenvalues. The curve at the last few factors/components lays low as scree and the values are regarded as random, or unaffected so much by the data.

Therefore visually one would look for the point at which the curve changes its slope notably (referred to as elbow) and consider only factors with corresponding eigenvalues that lay above this elbow on the plot. The rest is considered as random.

However this method is subjective, especially when there are more than one rapid changes in the curve’s slope. Also, this method cannot be applied in simulation studies as the current one, since this would require the visual interpretation of hundreds or thousands of plots. Therefore non-graphical ways for examining the sudden changes in the curve of the scree plot are developed (Raiche, Roipel, and Blais, 2006). Such methods are optimal coordinate and the acceleration factor, the latter examined in this thesis.

The acceleration factor finds the point on the graph, at which the curve changes most distinctively and considers only factors with eigenvalues bigger than those at the elbow point, which at the same time are bigger than 1 (or 0) with accordance to Kaiser criterion. The acceleration factor is the second derivative of the scree curve:

$$f''(i) = \frac{f(i + h) - 2f(i) + f(i - h)}{h^2},$$  \hspace{1cm} (18)

where $h$ is 1 and therefore:

$$f''(i) = f(i + 1) - 2f(i) + f(i - 1).$$  \hspace{1cm} (19)

More specifically, the one can detect the index of the factor/component at which the abrupt change in the curve is found:

$$\max_i(\lambda_{i+1} - 2\lambda_i + \lambda_{i-1}).$$  \hspace{1cm} (20)

The number of the components would be $i - 1$ as long as the $(i - 1)\text{th}$ eigenvalue satisfies the Kaiser criterion.
2.4.3. Parallel Analysis

The parallel analysis is suggested by Horn (1965) and it is also based on the scree test. The procedure generates a number of random matrices with dimensions corresponding to the original data and examines the scree plots obtained from their correlation matrices. The generated data is normally distributed. Additionally, re-sampling is performed from the original data using bootstrap and again the scree plots are examined. Typically, the results obtained from pure random data or from random data and bootstrap combined, don’t differ. In this thesis the combination of re-sampled and simulated data is used.

The eigenvalues of the correlation matrices for each generated dataset are calculated. The eigenvalues for the corresponding factors/ components are averaged and plotted against the indices of their factors. The curve for the eigenvalues obtained from the original data is also represented on the same plot. This will clearly show which eigenvalues of the original data are larger than or equal to those of the simulated data. It is precisely these eigenvalues, that are selected and the corresponding factors retained. The comparison of the eigenvalues, obtained from the original data and the simulated data, does not necessarily have to be graphical, but it is a widely applied practice.

3. Methodology: Design of the Monte Carlo simulation

3.1. Defining a set of data generating processes (DGP)

3.1.1. DGP used for detecting the number of principal components

Since the purpose of the thesis is to illustrate the accuracy of the described in the previous section methods in determining the number of components in PCA and factors in FA through simulations, data with specific characteristics must be generated multiple times for performing the analysis. Variables $X_1, X_2, ..., X_p$ must be generated, so that they can be reduced to a specific number of principal components and check if Kaiser criterion, acceleration factor and parallel analysis correctly identify this number. The question at hand is how to generate the data. In PCA, if the dataset of variables $X_1, X_2, ..., X_p$ is known, then the principal components can be computed according to the equation:

$$ Y_{n \times p} = X_{n \times p} W_{p \times p}, $$

where $Y_{n \times p}$ is the matrix of principal components, $X_{n \times p}$ is the data matrix and $W_{p \times p}$ is the matrix with columns – the eigenvectors, obtained from the data correlation matrix: $W \equiv [e_1 \ e_2 \ ... \ e_p]$.

Based on the equation above, in the DGP for $X_{n \times p}$ given $Y_{n \times p}$ and $W_{p \times p}$ can be deducted as follows:

$$ YW' = XWW' $$

$$ YW' = XI = X $$
The above is true, since matrix \( W \) is orthogonal and multiplied by its transpose gives the identity matrix.

However, the idea behind generating \( X \) is to be able to control the number of components, to which the data could be transformed to. To do that the number of columns of matrix \( W \) is reduced from \( p \) to \( k \). To compensate for the reduction, an error term is added. Therefore the DGP for PCA becomes:

\[
X_{n \times p} = Y_{n \times k}^* W_{k \times p}^* + E_{n \times p}.
\]

Note that matrix \( Y \) was reduced to having \( k \) columns and the transpose of matrix \( W \) now has \( k \) rows. The true number of principal component in the data is \( k \).

To generate the data matrix \( X \) however, first matrices \( Y_{n \times k}^* \), \( W_{k \times p}^* \) and \( E_{n \times p} \) must be generated.

The \( k \) vectors of principal component scores are generated randomly each from standard normal distribution \( N(0,1) \). Matrix \( E_{n \times p} \) has multivariate normal distribution as each column of the matrix is also generated from \( N(0,1) \).

To create the matrix of eigenvectors, first a correlation matrix \( R_{gen} \) must be generated. Joe (2006) suggests a procedure for generating correlation matrices given a particular partial correlation, which if set to 1 would give a positive definite matrix with diagonals 1 and correlation coefficients drawn from a \( Beta(p/2, p/2) \) distribution on \((-1, 1)\), or \(2 \cdot Beta(p/2, p/2) - 1\). Only matrices with at least \( k \) eigenvalues greater than 1 are chosen.

The eigenvectors of \( R_{gen} \) can be computed to form matrix \( W \), which is reduced to matrix \( W_{p \times k}^* \).

### 3.1.2. DGP used in detecting the number of factors

Using the same procedure as in the DGP for PCA to generate \( R_{gen} \), the data simulated for FA is obtained by using the following equation:

\[
X_{n \times p} = Y_{n \times k} L_{k \times p} + E_{n \times p},
\]

where

\[
L = \begin{bmatrix}
\sqrt{\lambda_1} e_1 \\
\sqrt{\lambda_2} e_2 \\
\vdots \\
\sqrt{\lambda_k} e_k
\end{bmatrix}
\]

and \( e_1, e_2, ..., e_{1k} \) are the eigenvectors of \( R_{gen} \) corresponding to the first \( k \) eigenvalues \( \lambda_i \).

Matrices \( Y_{n \times k} \) and \( E_{n \times p} \) are generated identically to the procedure for PCA.

From equations (22) and (23) one can see that the DGP for PCA and FA doesn’t differ much. In fact although the theory behind PCA and FA differs, giving different meaning to the weights in \( W_{k \times p}^* \) and \( L_{k \times p} \) as well as to the matrix \( E_{n \times p} \) in the two equations, the
fundamental process of generating the data is in fact the same. For the sole process of simulating the data, the true meaning of the elements on the right side of the equations are unimportant. Still, from the similarities between equations (22) and (23) one can infer that PCA is able to capture a factor structure of the data. What would differentiate the results obtained from PCA and FA for the performance of the Kaiser criterion, acceleration factor and parallel analysis is hidden in the estimation process and interpretation rather than the generation process. Nevertheless, to adhere to the theory behind PCA and FA, different weights as dictated by $W_{k \times p}^*$ and $L_{k \times p}$ are used for PCA and FA respectively as shown in equations (22) and (23).

### 3.2. Generating r independent replications

Each of the DGP’s is repeated $r$ times. For each replication different matrices $Y, L, W, E, X$ are generated. This would allow the screening of the accuracy of the methods for detecting the number of components and factors in each replication and a conclusion could be made on the basis of the overall performance of the methods with the help of a comprehensive measure for summarizing the results of all replications.

### 3.3. Estimations

The correlation matrix is used in this research in order to perform PCA and FA. For the factor analysis the correlation matrix is adjusted, so that it accounts only for the communalities in the variance. The elements on the diagonals are less than 1. The eigenvalues and eigenvectors of the corresponding correlation matrix are calculated.

The number of principal components selected by Kaiser criterion are determined by pinpointing the eigenvalues greater than 1. The number of factors is chosen by counting factors with eigenvalues greater than 0.

The acceleration factor shows the factor or component, at which a change on the scree curve is observed. The factors/ components above this elbow and eigenvalues greater than 1 (or 0) are chosen. Their number is saved.

The parallel analysis is performed with 100 iterations and uses both re-sampled and simulated data.

For measurement of accuracy, the mean absolute deviation from the true number $k$ of factors/ components is calculated:

$$MD = \frac{\sum_{i=1}^{r}|k_i^{detected} - k|}{r}. \quad (24)$$

The percentage of correctly identified number of factors/ components from the replications for each method is also calculated. Additionally, the average number of factors and components detected from all replications is computed, which helps determining the direction of the deviation – whether the method over- or underestimates $k$. A visual aid in this regard is the distribution of detected number of factors and components shown in bar
plots for each type of analysis (see Appendix). The mean deviations are also plotted for different values of $k, p$ and $n$.

4. Implementation

4.1. Data and model of the data generating process

The data generating process begins with simulating a correlation matrix $R_{gen}$ with correlation coefficients drawn from $2 \cdot \text{Beta}(p/2, p/2) - 1$ and $k$ eigenvalues greater than 1. The eigenvalues and eigenvectors of that matrix are found and matrices $W_{k \times p}^*$ and $L_{k \times p}$ computed.

Next, the principal component or factor scores are generated. If $Y_{n \times 1,j}^*$ represents the $j^{th}$ column of matrix $Y_{n \times k}^*$, then the DGP for the factors/components is:

$$Y_{n \times 1,j}^* \sim N(0,1), \quad j = 1, \ldots, k. \quad (25)$$

The entries in matrix $E_{n \times p}$ are also generated from standard normal distribution.

The dataset used for determining the number of principal components ($X_{n \times p}^{PC}$) and factors ($X_{n \times p}^{F}$) is computed as:

$$X_{n \times p}^{PC} = Y_{n \times k}^* W_{k \times p}^{*\prime} + E_{n \times p}; \quad (26)$$

$$X_{n \times p}^{F} = Y_{n \times k}^* L_{k \times p} + E_{n \times p}. \quad (27)$$

The simulations are performed with:

- $p = 4, 5$ and 10 number of variables $X$;
- $k = 1, 2$ and 3 true number of factors/components for $p = 4$
- $k = 1, 2, 3$ and 4 for $p = 5, 10$;
- $n = 100, 500$ and 1000 number of observations.

The number of replications for each combination of $p, k$ and $n$ is 1000. The estimations are done according to the described in the previous section methodology.

4.2. Results

4.2.1. Detected number of principal components

Table 1 below displays the mean absolute deviation from the true number of components, detected from using the acceleration factor, parallel analysis and Kaiser criterion. As expected with the increase in the sample size, the errors in detecting the number of components decrease.
It is easier to detect the differences in the performance of the methods using visual representation of Table 1. Plots of the mean deviations from the true number of components are presented in Figure 1 a) to l). It is evident that the most commonly used method – Kaiser criterion, accurately detects the number of components only for specific cases, when the true number of components is neither large or small considering the number of variables. Figure 1 g-i) shows this the most clearly - when there are 10 variables in the dataset, the accuracy of Kaiser criterion increases with the increase of k. When k = 1, Kaiser criterion would overestimate the number of components even for large samples, unless the number of variables is smaller in comparison with k. Kaiser criterion is most misleading for datasets, formed from small samples, large number of variables and small true number of components. From the tested combinations in the simulation, Kaiser criterion performs the worst for k = 1, p = 10 and n = 100. Table 2 showing the average number of components detected in the 1000 replication, reveals that in this case Kaiser criterion would detect an average of 4.269 components, instead of 1. In practice the true number of components is unknown and thus it would be hard to assess whether the use of Kaiser criterion would be appropriate, unless the number of variables is small, in which case it is more robust.

The acceleration factor, aiming to represent the scree test method, accurately estimates the true number of components for datasets with larger sample size and small true number of components in comparison to the number of variables. In fact for those cases, the acceleration factor shows the most accurate assessment, compared to the other methods, although the parallel analysis is almost as accurate. However, using the acceleration factor, the mean absolute deviation in Table 1 generally increases for larger number of components. The closer k gets to p, the worst the performance of the acceleration factor. Table 2 shows that the acceleration factor would give completely erroneous results for small number of variables and larger true number of components (p = 4, k = 3 and p = 5, k = 4). In this case even for large samples, the acceleration factor didn’t manage to correctly detect the number of components in any of the 1000 replications (Table 3). Although the big ratio k:p affects the other two methods too, the acceleration factor is the most affected method of those 3. Figure 1 g) to l) shows that the sample size affects the acceleration factor, along with the other methods and is more accurate for larger n.
Figure 1: Mean absolute deviation of detected number of components from the true number of components \((k)\) for \(p = 4, 5\) and 10 number of variables and sample size 100, 500 and 1000

a) \(p = 4, n = 100\)

b) \(p = 4, n = 500\)

c) \(p = 4, n = 1000\)
d) $p = 5, n = 100$

e) $p = 5, n = 500$

f) $p = 5, n = 1000$

g) $p = 10, n = 100$

h) $p = 10, n = 500$

i) $p = 10, n = 1000$
Unlike the Kaiser criterion or the acceleration factor, the parallel analysis doesn’t over- or underestimate the true number of components in an extreme way in any of the combinations of \( k, p \) and \( n \) used in the simulation. For larger sample sizes like \( n=500 \) or \( n=1000 \), this method gives accurate estimations in at least 95% of the replications. Even for smaller sample sizes like \( n=100 \), the parallel analysis leads to small mean absolute deviation from the true number of components.

Although all of the three methods are affected by the ratio \( k: p \) and perform poorly for ratio closer to 1, it can be said that overall the acceleration factor performs worse for larger number of components and smaller number of variables contrary to Kaiser criterion, which exhibits larger mean absolute deviation from the true number of components for small \( k \) and larger number of variables. The parallel analysis however, performs well regardless of the true number of components and number of variables and is affected only by the number of observations, but not severely. Still, the acceleration factors can outperform the parallel analysis for \( k = 1 \).

**Table 2: Average number of components detected in 1000 replication for \( k = 1, 2 \) and 3 true number of components, \( p = 4, 5 \) and 10 number of variables and sample size 100, 500 and 1000**

<table>
<thead>
<tr>
<th>Sample Size (n)</th>
<th>Method</th>
<th>( p = 4 )</th>
<th>( p = 5 )</th>
<th>( p = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k = 1 )</td>
<td>( k = 2 )</td>
<td>( k = 3 )</td>
<td>( k = 1 )</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>1.015 1.976 1.002 1.004 2.000 2.958 1.372 1.001 2.000 2.994 3.987</td>
<td>1.037 1.991 2.944 1.051 2.003 2.988 3.876 1.286 2.040 3.001 3.999</td>
<td>1.192 2.000 2.788 1.455 2.002 2.986 3.368 3.677 2.956 3.061 4.000</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>1.000 1.991 1.000 1.000 2.001 2.990 1.352 1.000 2.000 3.001 3.997</td>
<td>1.017 1.998 2.980 1.030 2.001 2.998 3.972 1.116 2.008 3.000 4.000</td>
<td>1.132 2.000 2.913 1.358 2.001 2.998 3.624 3.336 2.490 3.010 4.000</td>
</tr>
<tr>
<td>500</td>
<td>Acceleration Factor</td>
<td>1.015 1.976 1.002 1.004 2.000 2.958 1.372 1.001 2.000 2.994 3.987</td>
<td>1.037 1.991 2.944 1.051 2.003 2.988 3.876 1.286 2.040 3.001 3.999</td>
<td>1.192 2.000 2.788 1.455 2.002 2.986 3.368 3.677 2.956 3.061 4.000</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>1.015 1.976 1.002 1.004 2.000 2.958 1.372 1.001 2.000 2.994 3.987</td>
<td>1.037 1.991 2.944 1.051 2.003 2.988 3.876 1.286 2.040 3.001 3.999</td>
<td>1.192 2.000 2.788 1.455 2.002 2.986 3.368 3.677 2.956 3.061 4.000</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>1.000 1.991 1.000 1.000 2.001 2.990 1.352 1.000 2.000 3.001 3.997</td>
<td>1.017 1.998 2.980 1.030 2.001 2.998 3.972 1.116 2.008 3.000 4.000</td>
<td>1.132 2.000 2.913 1.358 2.001 2.998 3.624 3.336 2.490 3.010 4.000</td>
</tr>
<tr>
<td>1000</td>
<td>Acceleration Factor</td>
<td>1.015 1.976 1.002 1.004 2.000 2.958 1.372 1.001 2.000 2.994 3.987</td>
<td>1.037 1.991 2.944 1.051 2.003 2.988 3.876 1.286 2.040 3.001 3.999</td>
<td>1.192 2.000 2.788 1.455 2.002 2.986 3.368 3.677 2.956 3.061 4.000</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>1.015 1.976 1.002 1.004 2.000 2.958 1.372 1.001 2.000 2.994 3.987</td>
<td>1.037 1.991 2.944 1.051 2.003 2.988 3.876 1.286 2.040 3.001 3.999</td>
<td>1.192 2.000 2.788 1.455 2.002 2.986 3.368 3.677 2.956 3.061 4.000</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>1.000 1.991 1.000 1.000 2.001 2.990 1.352 1.000 2.000 3.001 3.997</td>
<td>1.017 1.998 2.980 1.030 2.001 2.998 3.972 1.116 2.008 3.000 4.000</td>
<td>1.132 2.000 2.913 1.358 2.001 2.998 3.624 3.336 2.490 3.010 4.000</td>
</tr>
</tbody>
</table>

The mean absolute deviation does not show the direction of deviation from the true number of components. **Table 2** is useful in that it shows whether the methods over- or underestimate that number. When there is only one component, naturally the methods can only overestimate \( k \). Wherever, the acceleration factor proved being inaccurate, it underestimated the true number of components. That is, for \( k = 3 \), it detected an average of 1.125 components for \( n = 100 \) and \( p = 4 \); an average of 1.002 components for \( n = 500 \) and \( p = 4 \) and an average of 1.000 component for \( n = 1000 \) and \( p = 4 \). For 100 observations, 5 variables and 4 components, the acceleration factor detected an average of 1.536 components in the 1000 replications. For 10 variables in the dataset of 100 observations, the mean number of components detected by the acceleration factor is 2.483 for \( k = 3 \) and 2.836 for \( k = 4 \) (also see Figure 9 bb and ee in Appendix). Kaiser criterion generally overestimates the true number of components and extremely does so for cases with smaller \( k \) and larger number of variables. It is possible however for the Kaiser criterion to underestimate \( k \) if the ratio \( k:p \) is close to 1. Similarly to the
acceleration factor, the Kaiser criterion is affected by very large true number of components. (also see Figure 9 s in Appendix).

Table 3: Percentage (%) of correctly detected number of components from 1000 replication for \( k = 1, 2 \) and 3 true number of components, \( p = 4, 5 \) and 10 number of variables and sample size 100, 500 and 1000

<table>
<thead>
<tr>
<th>Sample Size (n)</th>
<th>Method</th>
<th>( p = 4 )</th>
<th>( p = 5 )</th>
<th>( p = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k = 1 )</td>
<td>( k = 2 )</td>
<td>( k = 3 )</td>
<td>( k = 1 )</td>
</tr>
<tr>
<td>100</td>
<td>Acceleration Factor</td>
<td>87.7</td>
<td>81.3</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>81.9</td>
<td>88.8</td>
<td>53.4</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>45.8</td>
<td>98.1</td>
<td>25.4</td>
</tr>
<tr>
<td>500</td>
<td>Acceleration Factor</td>
<td>98.5</td>
<td>97.6</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>96.3</td>
<td>98.9</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>80.8</td>
<td>100.0</td>
<td>78.8</td>
</tr>
<tr>
<td>1000</td>
<td>Acceleration Factor</td>
<td>100.0</td>
<td>99.1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>98.3</td>
<td>99.8</td>
<td>98.0</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>86.8</td>
<td>100.0</td>
<td>91.3</td>
</tr>
</tbody>
</table>

Table 3, which shows the percentage of correctly detected number of components, gives more specific information on the accuracy of the 3 methods. It can be seen for example that in none of the 1000 replication did the Kaiser criterion correctly identify the number of components for datasets having 10 variables and 1 underlying component. This highlights the extreme overestimation of the criterion in those cases. In general, Kaiser criterion has poor performance for datasets with sample size of 100. Another extreme case as already mentioned is the acceleration factor, which failed to detect \( k = 3 \) in cases, where the number of variables is 4 and \( k = 4 \) in cases when the number of variables is 5. Although the acceleration factor has much larger mean absolute deviation for 5 variables, 4 true number of components and 100 observations, and Kaiser criterion didn’t underestimate the number of components as severely as the acceleration factor, Table 3 shows that it still failed to correctly detect the 4 components in any of the 1000 replications. Instead, it detected an average of 2.732 components.

The parallel analysis however performs well and for samples of 500 observations and more, it has more than 95% accuracy in almost all generated datasets. Exclusion is the case with 10 variables and 1 component, where the method accurately estimated \( k \) is 89.4% and 75.7% of the cases accordingly for \( n = 1000 \) and \( n = 500 \). Also, for \( k = 4 \) and \( p = 5 \), the accuracy is 88.8%. For sample size of 100 observations, overall the method is more robust than Kaiser criterion or parallel analysis.

4.2.2. Detected number of factors

The results, obtained from simulating data for factor analysis are similar to those for principal component analysis. The mean absolute deviation from the true number of factors is presented in Table 4. Again by increasing the sample size the accuracy of the methods improves.
Table 4: Mean absolute deviation of detected number of factors from the true number of factors ($k$) for $p = 4, 5$ and $10$ number of variables and sample size $100, 500$ and $1000$

<table>
<thead>
<tr>
<th>Sample Size (n)</th>
<th>Method</th>
<th>$p = 4$</th>
<th>$p = 5$</th>
<th>$p = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k = 1$</td>
<td>$k = 2$</td>
<td>$k = 3$</td>
</tr>
<tr>
<td>100</td>
<td>Acceleration Factor</td>
<td>0.011</td>
<td>0.245</td>
<td>1.824</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>0.231</td>
<td>0.050</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>0.642</td>
<td>0.003</td>
<td>0.351</td>
</tr>
<tr>
<td>500</td>
<td>Acceleration Factor</td>
<td>0.000</td>
<td>0.164</td>
<td>1.989</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>0.005</td>
<td>0.003</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>0.253</td>
<td>0.000</td>
<td>0.057</td>
</tr>
<tr>
<td>1000</td>
<td>Acceleration Factor</td>
<td>0.000</td>
<td>0.130</td>
<td>1.997</td>
</tr>
<tr>
<td></td>
<td>Parallel Analysis</td>
<td>0.029</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Kaiser Criterion</td>
<td>0.193</td>
<td>0.000</td>
<td>0.014</td>
</tr>
</tbody>
</table>

The results from Table 4 are also illustrated on Figure 2 (a) to (i). For the factors, as for the principal components, Kaiser criterion is accurate only in cases, for which the true number of factors is relatively large with respect to the number of variables, but not too close to $p$. Kaiser criterion’s estimates are extremely overestimated for cases, where $k = 1$. Especially large is the bias for smaller sample sizes and the mean absolute deviation for $n = 100$ and $p = 10$ is $5.182$, $4.926$ and $4.764$ for $k = 1$, $k = 2$ and $k = 3$ accordingly. The performance of Kaiser criterion aggravates with the increase of the number of variables in the datasets. In practice, when exploratory factor analysis is performed, the true number of factors is unknown and there is no way for evaluating the usefulness of Kaiser criterion for a specific dataset. It’s therefore only appropriate to use Kaiser criterion for sets with small number of variables.

The estimates, obtained by using the acceleration factor are accurate only for datasets with small number of underlying factors – just contrary to Kaiser criterion. Again, in practice it would be hard to justify the use of this method as the true number of factors is unknown. The mean absolute deviation, presented in Table 4 increases with the increase of $k$ and get to extreme values if the ratio $k:p$ gets close to 1. Parallel analysis is superior to the acceleration factor for large number of factors. The acceleration factor improves with the increase of the number of variables.

Judging by Figure 2, it seems that the parallel analysis has the overall smallest mean average deviations from the true number of factors. It is also not affected by the number of factors. Naturally, it improves by increasing the sample size, but gives satisfactory results even for samples of size 100.
Figure 2: Mean absolute deviation of detected number of factors from the true number of factors \((k)\) for \(p = 4, 5\) and 10 number of variables and sample size 100, 500 and 1000

a) \(p = 4, n = 100\)  
b) \(p = 4, n = 500\)  
c) \(p = 4, n = 1000\)
The parallel analysis performs well and its smallest percentage of correctly detected number of factors is 71.4% for 100 observations, 5 variables and 1 factor. For the same dataset
Kaiser criterion detects correctly $k$ in only 3.5% of the 1000 replications. The Acceleration factor showed the best results in that particular dataset as it performs well for small number of factors. Increasing the factors however, the parallel analysis is superior over Kaiser criterion or the acceleration factor.

4.3. Example “Places” dataset

An example dataset has been used in order to demonstrate the use and performance of the studied methods in practice. The number of components and factors are determined based on the findings from the Monte Carlo simulations.

The dataset is taken from “Places Rated Almanac”, by Richard Boyer and David Savageau and published by Rand McNally. The dataset rates 329 communities in USA, based on 9 characteristics:

1) Climate and Terrain  
2) Housing  
3) Health Care & the Environment  
4) Crime  
5) Transportation  
6) Education  
7) The Arts  
8) Recreation  
9) Economics  

The data is accessible in R using package `tourrr` under command ‘places’.

4.3.1. Number of principal components

Suppose that we had only the first 4 variables – the ratings for climate and terrain, housing, health care and environment, and crime, and apply the methods for choosing the number of components to those variables only.

The correlation matrix for the extracted dataset is:

$$R = \begin{bmatrix} 1.00 & 0.39 & 0.21 & 0.19 \\ 0.39 & 1.00 & 0.45 & 0.13 \\ 0.21 & 0.45 & 1.00 & 0.30 \\ 0.19 & 0.13 & 0.30 & 1.00 \end{bmatrix}$$

The matrix has eigenvalues: $\lambda = [1.861 \ 0.906 \ 0.773 \ 0.460]$. According to Kaiser criterion only one component should be extracted as there is only one eigenvalue greater than 1. Figure 3 shows the scree plot. The scree test would suggest using only 1 component as the elbow of the curve corresponds to the second component. The acceleration factor scree test confirms that as the acceleration factors for the second and third components are:

$$af_2 = \lambda_3 - 2 \lambda_2 + \lambda_1 = 0.773 - 2 \cdot 0.906 + 1.861 = 0.822$$

$$af_3 = \lambda_4 - 2 \lambda_3 + \lambda_2 = 0.460 - 2 \cdot 0.773 + 0.906 = -0.180$$
The elbow is in the second component since $a_{f_2} > a_{f_3}$.

Figure 3: Scree plot for PCA for “Places” dataset with the first 4 variables used (n = 329)
The dotted curve shows the mean eigenvalues, obtained from simulated data in parallel analysis

The parallel analysis also points out to only 1 components as there is only one eigenvalue greater than the mean eigenvalue curve, obtained from the simulated data.

For 4 variables, all three methods used in principal component analysis for choosing the number of components indicated that $k = 1$ - the 4 variables can be outlined by one principal component only. This is consistent with the findings from the performed simulations, since for samples of 500 observations and 1 underlying component, in 96.3% of the replications the acceleration factor correctly detected that $k = 1$, for parallel analysis – it’s 80.8%, and for Kaiser criterion – 100%. The sample size in this example is less than 500 observations (it’s 329) and although the acceleration factor and the parallel analysis don’t worsen so notably for sample of 100 in the performed simulations, Kaiser criterion does. Still one can deduce that 329 observations were enough for Kaiser criterion to identify the correct number of components.

Next one more variable is added to the analysis. Figure 4 shows the scree plot for 5 variables. There are two eigenvalues greater than 1, so Kaiser criterion chooses 2 components. The acceleration factor chooses 1 component, since the elbow is in the second, and the parallel analysis also chooses 1 component, since only one eigenvalue lies above the simulation eigenvalue curve.
Most probably, the true number of components in this case is indeed 1. The second eigenvalue of 1.02 barely paces the barrier of 1. According to the performed Monte Carlo study for $p = 5$ and $k = 1$ and $n = 500$, in 99.6% of the cases the acceleration factor correctly identifies that there is only one component. For parallel analysis this percentage is 94.9% and for Kaiser criterion – 54.8%, just above half of the replications. For 100 observations, the acceleration factor has 84.5% chance of correctly detecting $k=1$, parallel analysis – 72.5%, but Kaiser criterion’s performance worsens remarkably as only in 12.9% of the replications did it correctly identify that there is 1 component. Therefore, one would expect that for 329 observations, Kaiser criterion would have less than 50% chance of correctly identifying the true number of components in this case.

Next, the methods are applied to the full dataset of 9 variables. The scree plot is shown in Figure 5. The eigenvalues are:

$$\lambda = [3.408 \ 1.214 \ 1.141 \ 0.921 \ 0.753 \ 0.631 \ 0.493 \ 0.318 \ 0.120].$$

Kaiser criterion would select 3 components. The elbow is in the second component, indicating that the scores of only one component should be estimated. The mean eigenvalues from the simulated data in the parallel analysis is:

$$\overline{\lambda}_{simulated} = [1.266 \ 1.175 \ 1.104 \ 1.049 \ 0.993 \ 0.940 \ 0.884 \ 0.828 \ 0.760].$$

The first three values in $\lambda$ are greater than those in $\overline{\lambda}_{simulated}$ indicating that 3 components should be chosen according to the parallel analysis with the third eigenvalue being barely above the third simulated value. What’s the true number of components?
If the true number of components is 1, then it is possible to obtain result as the above. The acceleration factor would correctly identify that there is one underlying component in 99.9% of the cases for samples with size 500 and in 74.7%, when the size is 100 observations, when the number of variables is 10. For small number of components, Kaiser criterion highly overestimates the number of components in all the replications in the simulation. Parallel analysis is not reliable, when the true number of components is 1. For sample size of 100 observations, the chances of it overestimating the number of components is around 50% with about 25% chances of the estimate being 3 or above (see Figure 9 v) – w) in Appendix). Although 2 of the three methods point to 3 components, judging by the results, obtained in the Monte Carlo simulations, it seems probable that the true number of components is just 1.

4.3.2. Number of factors

The same dataset can be used for identifying latent variables and group the original variables into groups, affected by the same factors. If the analyzed methods for choosing the number of factors are applied to the “Places” dataset restricted to only the first 4 variables, the adjusted correlation matrix with the estimations of the communalities replacing the ones on the diagonal, would be:

$$ R_{adj} = \begin{bmatrix} 0.17 & 0.39 & 0.21 & 0.19 \\ 0.39 & 0.30 & 0.45 & 0.13 \\ 0.21 & 0.45 & 0.27 & 0.30 \\ 0.19 & 0.13 & 0.30 & 0.11 \end{bmatrix} $$
The eigenvalues are: $\lambda = [1.096 \quad 0.070 \quad -0.028 \quad -0.295]$. According to Kaiser criterion two factors should be extracted as there are two eigenvalues greater than 0. Figure 6 shows the scree plot. The scree test suggests 1 factor as the elbow of the curve corresponds to the second factor ($af_2 > af_3$):

$$af_2 = \lambda_3 - 2\lambda_2 + \lambda_1 = 0.928$$
$$af_3 = \lambda_4 - 2\lambda_3 + \lambda_2 = -0.169.$$ 

The mean eigenvalues of the simulated data are:

$$\tilde{\lambda}_{\text{simulated}} = [0.135 \quad 0.039 \quad -0.131 \quad -0.107].$$

Three of the eigenvalues are greater than the simulated ones, suggesting the use of 3 factors.

*Figure 6: Scree plot for FA for “Places” dataset with the first 4 variables used (n = 329)*

The dotted curve shows the mean eigenvalues, obtained from simulated data in parallel analysis.

The three methods suggest different number of factors. The acceleration factor recommends using 1 factor, the parallel analysis – 3, and Kaiser criterion – 2. In the simulated data, similar results can be obtained when the true number of factors is 3: for 4 variables and $k=3$, the acceleration factor failed to detect the number of factors in all 1000 replications suggesting an average of 1.176 factors, when using a sample of size 100 and 1.011 factors for sample size of 500 observations. Kaiser criterion also tends to underestimate the number of factors for samples smaller than 500 observations. One can conclude that the number of factors is indeed 3, when the first 4 variables of the “Places” dataset are used.

Let’s add the fifth variable to the dataset. The scree plot is illustrated on Figure 7.
Figure 7: Scree plot for FA for “Places” dataset with the first 5 variables used (n = 329)
The dotted curve shows the mean eigenvalues, obtained from simulated data in parallel analysis

Parallel Analysis Scree Plots

Acceleration factor suggests 1 factor, parallel analysis – 3 and Kaiser criterion also 3. This results correspond to a true number of factors being 3 as for 5 variables Parallel analysis and Kaiser criterion can detect that \( k = 3 \) with 92.4% and 96.8% accuracy respectively for sample size 100 and 99.6% and 99.8% for sample size 500. The acceleration factor would underestimate the true number of factors being 3 in 40% of the cases with around 30% chance of suggesting 1 factor instead of 3, when the sample size is 100. Given that \( k=3 \) and the sample size is 500, acceleration factor would underestimate the number of factors in about 30%, suggesting 1 factor in above 20% of the replications in the Monte Carlo study (see Figure 10 p) – q) in Appendix). All this leads to the conclusion that the true number of factors in this case is 3.

Using the full “Places” dataset gives the following scree plot – Figure 8.
Acceleration factor suggests using 1 factor, Parallel and Kaiser criterion – 6. No simulations were run with 6 factors, but from the simulations run with 10 variables, one can assume that the true number of factors is in fact 6, because of the following:

- If the true number of factors were 1, based on the simulations one would expect that the Kaiser criterion would overestimate k, which is consistent to the example (Kaiser criterion suggests 6 factors), but the parallel analysis couldn’t have been so wrong as to show 6 factors instead of 1. Parallel analysis doesn’t overestimate the true number of factors so greatly, when $k=1$.

- Increasing the number of variables when 10 variables were used in the simulations lead to more robust Kaiser criterion and parallel analysis, but worsened the acceleration factor, which started underestimating the number of factor with the increase of $k$.

- In the previous two examples, when less variables were used, 3 factors were indicated. Having less than 3 factors, when the number of variables increases would be inconsistent with the previous results.

All this leads to concluding that the number of factors is 6.
5. Summary and Conclusion

5.1. Summary

A Monte Carlo study has been performed for assessing the accuracy of three frequently
used methods for detecting the number of factors and components in factor analysis and
principal component analysis: Kaiser criterion, acceleration factor and parallel analysis.

Data was generated independently in a sequence of 1000 replications for principal
component analysis and factor analysis, so that it would have a specific number of
underlying components or factors: $k$. Then the three methods for detecting the number of
factors/ components were applied and their overall performance examined. The mean
absolute deviation from the true number $k$ was computed as well as the average number of
detected factors/ components and the percentage of correct estimates from all 1000
replications.

The procedure was performed by simulating datasets with 4, 5 and 10 number of variables,
having from 1 to 4 true number of factors/ components ($k = 4$ was tested, when 10 and 5
variables were simulated). The datasets were generated with sample size of 100, 500 and
1000 for all combinations of $k$ and $p$ (number of variables).

5.2. Conclusion

The following conclusions can be drawn from the results, obtained by the Monte Carlo
simulations:

- By increasing the number of observations in the datasets, the accuracy of all three
methods improves both for detecting the number of components and factors.
- Kaiser criterion is accurate only when the true number of components or factors is
neither small or large with respect to the total number of variables. Kaiser criterion
shows worse results, when there is only 1 underlying component.
- Kaiser criterion overestimates extremely the number of factors/ components when
the $k$ is smaller and the number of variables relatively large. In some cases Kaiser
would not recognize the true number of factors/ components in any of the
performed 1000 replications.
- Although Kaiser criterion generally overestimates the number of components, it can
also underestimate their number in cases when the true number of components is
closer to the number of variables.
- The acceleration factor performs well for larger sample sizes unless the true number
of factors/ components is close to the number of variables. It would therefore have a
better chance of correctly identifying the number of factors/ components in cases
when a lot of variables are used.
- When the sample size is not sufficiently large the acceleration factor tends to
underestimate the number of factors/ components. It performs well only for small $k$. 

- The acceleration factor has bad performance when the true number of factors/components is larger and the number of variables – smaller. This is more severe with factor analysis.
- For factor analysis, the acceleration factor performs poorly when the true number of factors is large, even for larger samples and many variables.
- Parallel analysis doesn’t show extreme over- or underestimation of the number of factors/components in any of the simulated datasets. It is less accurate for smaller samples, but still performs better than Kaiser criterion or the acceleration factor.
- The acceleration factor outperforms the parallel analysis only for very small number of factors/components, which in practice are unknown.

Considering the above conclusions, it would be natural to recommend the use of parallel analysis over Kaiser criterion or acceleration factor. Parallel analysis generally outperforms Kaiser criterion. It is surpassed by the acceleration factor only, when the number of factors/components is small as evident in one of the examples. However, the true number of components is unknown in reality. Therefore, what can be done in practice is to try several methods for estimating \( k \) and knowing the weaknesses and strengths of all the methods, to assume what the true number of factors/components is. Something similar was done in the example section.

6. Recommendations

This thesis only analyses three of the most commonly used methods for selecting the number of factors and components in FA and PCA. However, there are other developed methods such as choosing the number of factors/components, so that combined they explain a particular ratio of the total variance (e.g. 80%), Bartlett Chi-squared test (Bartlett, 1950) and also the Minimum Average Partial procedure, developed by Velicer (1976) and used only in principal component analysis. A review of these methods would be a suitable continuation of the current thesis. Also, an Akaike or Bayesian information criterion could be suggested and tested.

As suggested in the previous section, the knowledge of studies, such as this one can help researchers pinpoint the correct number of factors or components, even if all the methods have given different estimates. To do so, one has to know the strengths and weaknesses of each method and how they compare to each other for different combinations of \( k, p \) and \( n \). However, computing all the three estimates and additionally validating the results, based on simulation studies similar to the one in this thesis would be computationally and time demanding. Perhaps, an algorithm can be suggested that does this automatically after the computation of the estimates from the classical methods, by assuming different options for the true number of factors and components and examining the probability of the 3 methods giving false or correct results in the exact combination in which they did. The probability of underestimating or overestimating \( k \) in each method given the assumed true number of factors/components must be considered. To formulate an algorithm such as this one however, a huge simulation has to be done beforehand including a lot of different combinations of \( k, p \) and \( n \), a lot more than those done in the current study.
7. Bibliography


8. Appendix

8.1. Frequency plots

8.1.1. Detected number of components in all 1000 replications
Figure 9: Frequency plot of the detected number of components in all 1000 replications

a) $p = 4, k = 1, n = 100$
100 observations, 1 true component and 4 variables

b) $p = 4, k = 1, n = 500$
500 observations, 1 true component and 4 variables

c) $p = 4, k = 1, n = 1000$
1000 observations, 1 true component and 4 variables

d) $p = 4, k = 2, n = 100$
100 observations, 2 true components and 4 variables

e) $p = 4, k = 2, n = 500$
500 observations, 2 true components and 4 variables

f) $p = 4, k = 2, n = 1000$
1000 observations, 2 true components and 4 variables
y) $p = 10, k = 2, n = 100$

100 observations, 2 true components and 10 variables
Detected number of components

bb) $p = 10, k = 3, n = 100$

100 observations, 3 true components and 10 variables
Detected number of components

d) $p = 10, k = 3, n = 1000$

1000 observations, 3 true components and 10 variables
Detected number of components

z) $p = 10, k = 2, n = 500$

500 observations, 2 true components and 10 variables
Detected number of components

cc) $p = 10, k = 3, n = 500$

500 observations, 3 true components and 10 variables
Detected number of components

aa) $p = 10, k = 2, n = 1000$

1000 observations, 2 true components and 10 variables
Detected number of components

dd) $p = 10, k = 3, n = 1000$

1000 observations, 3 true components and 10 variables
Detected number of components
ee) $p = 10, k = 4, n = 100$

100 observations, 4 true components and 10 variables

ff) $p = 10, k = 4, n = 500$

500 observations, 4 true components and 10 variables

gg) $p = 10, k = 4, n = 1000$

1000 observations, 4 true component and 10 variables
8.1.2. Detected number of factors in all 1000 replications

Figure 9: Frequency plot of the detected number of factors in all 1000 replications

a) $p = 4, k = 1, n = 100$

100 observations, 1 true factor and 4 variables

b) $p = 4, k = 1, n = 500$

500 observations, 1 true factor and 4 variables
c) $p = 4, k = 1, n = 1000$

1000 observations, 1 true factor and 4 variables