Universiteit
Leiden
The Netherlands

Latent Dimension Selection in High Dimensional Datasets

Determine the number of latent factors in high-dimensional factor analysis

Xiaochun Shao

Daily advisor: Dr.Carel F.W.Peeters

Biometris, Wageningen University & Research

Second advisor: Dr.Willem Kruijer

Biometris, Wageningen University & Research

Defended on Date , Year

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

Foreword

In this paper, we explore a new approach to determine the optimal number of latent dimensions in high dimensional data sets. When under the condition of the number of column dimension exceeding sample sizes, widely used dimension selection methods suffer from either over identification of latent dimensions or complicate to apply. Marchenko-Pastur law overcomes these kinds of problems. It states that the average distribution of eigenvalues from a random correlation matrix converges to a constant ratio of the number of column dimensions to sample sizes, p/n when $p \to \infty$.

We thank Dr.Carel.Peeters from Mathematical & Statistical Methods group (Biometris), Wageningen University & Research, for proposing this idea of implementing Marchenko-Pastur law to select optimal latent dimensions that originates from random matrix theory and having related theories proved.

Contents

Abstract

Background: Factor analysis has been an essential workhorse in analyzing high dimensional data sets, especially when the number of observed dimensions exceeds sample sizes. Determining the number of latent dimensions has an imperative impact on the quality and accuracy of factor models. Some existing methods of selecting the optimal number of latent factors, for instance, Kaiser's eigenvalue greater than one rule, maximum log likelihood ratio test, information criterion, scree plot test, and parallel analysis, mostly suffer from over identification of latent dimensions, which inevitably causes problems of inaccuracies in factor analysis.

Proposed method: We propose a method on the basis of Marchenko-Pastur law, which originates from random matrix theory. This law states that for a $n \times p$ random matrix following Wishart distribution,the limiting behaviour of its eigenvalues is generally bound by a constant ration $n/p \to c \in (0, \infty)$ as p approaches to ∞. The latent factors are selected as its respective eigenvalues greater that the value of bounding ratio of $c_+ = \left(1 + \sqrt{\frac{1}{c}}\right)^2$.

Results: In a simulation study, comparing with methods of Kaiser's rule, Empirical Kaiser Criterion, Eigenvalue ratio and Growth ratio, Marchenko-Pastur law performs the best in most of cases, particularly when the number of observed variables exceeds the sample sizes.

Keywords: Factor analysis; High dimensional data; Latent factors/dimensions; Marchenko-Pastur law

1 Introduction

1.1 Concept of factor analysis

Factor analysis has been an essential workhorse in many fields of sciences, such as biology, psychology, sociology and economics. For instance, psychologists and sociologists often encounter problems with numerous variables, for which they have to search for regularities in the behavioural characteristics of humans (Child, 2006). These "regularities" are inherent generalized construct that cannot be observed or measured explicitly. The core idea of factor analysis is hence a mathematical technique to identify underlying dimension of a construct among observed variables or measures that are interrelated with each other. Those identified "dimensions" are also regarded as factors (Pett et al., 2003). Dating back to the nineteenth and early twentieth centuries, a scientist named Galton laid the foundations of factor analysis. One of his major contributions in factor study was the concept of correlations, for which he attempted to use mathematical tools to formulate some ideas about the interdependence between two variables (Child, 2006). In the modern form of factor model study, the number of observed variables are expanded to exceed the number of subjects in most of cases. Nevertheless, the main idea is still more or less the same. Factor analytic model assumes that the observed variables can be represented by a lower-dimensional construct, which also denotes a smaller number of latent common factors. In other words, the observed variables are linear combinations of some unobserved factors, which generalizes the "true" meaning of observed variables (Kim et al., 1978).

Factor analysis basically has two different classes: Exploratory factor analysis (EFA) and Confirmatory factor analysis (CFA). The focus of exploratory factor analysis (EFA) is to explore a smaller set of latent common factors. Confirmatory factor analysis (CFA) is instead used to test for the hypothesis regarding the defined factor structure and what factors should be in the model (Henson and Roberts, 2006). To this end, exploratory factor analysis (EFA) leads to an inevitable problem that how to select the number of dimensions for which projecting the original data onto a lower dimensional space. In other words, the method to determination of the number of latent factors is the core question for our research.

1.2 Existing methods of determining latent dimensions

Typically, one of the most critical decisions in factor analysis is determining the optimal dimensions of projected space, in other words, the optimal number of latent factors, especially in the case of extremely high dimensions. Based on previous studies, a great deal of approaches have been proposed and implemented. One of the most widely used method is eigenvalue greater than one rule, which states that the composites with an eigenvalue larger than one are selected as one retained projected dimension (Kaiser, 1960). Guttman also suggested this rule to set the lower bound for common factors in a correlation matrix (Guttman, 1954). The problem of this easy-to-apply criterion is that it can easily lead to

over-extraction of the number of factors and hence poor accuracy (Velicer et al., 2000; Zwick and Velicer, 1986). Yet it is still largely employed by most computer software due to its simplicity.

Other than the universal use of Kaiser's rule for factor analysis, maximum log likelihood ratio test is the most formal method to determining the optimal number of latent factors retained(Peeters et al., 2019). This method employs hypothesis testing for which the alternative hypothesis is the saturated model with all the observed variables, whereas the null hypothesis is the model with parameters only representing the selected latent factors. And the test statistics follows a Chi-square distribution with the degrees of freedom as the difference in the number of factors between null and alternative model. However, a couple of concerns exist with respect to the maximum likelihood ratio test. Firstly, this method implies a so-called "badness-of-fit" test where it prefers non-significance in the test statistics (Velicer et al., 2000). This counter-intuitive preference can be achieved by purposely setting the power smaller. Moreover, it has the tendency to overly identify latent factors than actually implied in the data. The reason is that an eigenvalue representing only one observed variable might be retained as one of the common factors, which have shared variances across multiple observed variables (Velicer et al., 2000).

Determining the latent dimensions in factor analysis can also be treated as a model selection process. Information criterion are suitable methods for finding the optimal number of latent factors and simultaneously balancing model complexity (Peeters et al., 2019). These criterion calculate minus 2 times the latent factor dependent maximum log likelihood values plus a penalty of free parameters in the model. Akaike Information criterion (AIC) and Bayesian Information Criterion (BIC) are two most commonly used criterion. The procedure is to calculate AIC and BIC for a range of the number of latent factors. The factor model with lowest AIC and BIC scores is selected. This method often leads to overly identification of factors.

Scree plot also utilizes eigenvalues to determine the number of factors. Specifically, scree plot is a line drawn by eigenvalues on the y-axis against their respect ranks on the x-axis (Cattell, 1966). Its difference from Kaiser's rule is the cut-off value, which is a turning point that levels off the slope of the plotted line (Velicer et al., 2000). The number of eigenvalues above the turning point determines the optimal number of latent factors. The problem of subjectivity in eyeballing the turning point when the turning corner is not so obvious or when there are more than one turning points along the plotted lines (Velicer et al., 2000). Relying on visual identification, this method is proved to be accurate with large sample size and strong composites (Zwick and Velicer, 1986), and less accurate with low communalities and less sample sizes (Hakstian et al., 1982).

Besides those easy-to-apply approaches, simulation approaches have also been studied and implemented. For instance, parallel analysis is one of the simulation approaches introduced by Horn (Horn, 1965). Basically, when employing this method, many random correlation data matrices with the same number of variables and subjects as in the observed

data are simulated. Then eigenvalues are extracted for each random correlation matrix. The average of all eigenvalues across all simulated random correlation matrices are subsequently calculated. If the extracted eigenvalues in the observed data matrix exceeds the average eigenvalues, the aspect factors are retained as latent common factors. One of the problems arise from this method is how many random correlation matrices to generate to obtain a set of representative averaged eigenvalues. In the original study, only one random correlation matrix is generated, nevertheless, the number of generated random correlation matrices ought to be large enough to draw a stable curve (Horn, 1965). A large amount of computing power is thus required when the observed data is in high dimensions and the number of simulated correlation matrices is large.

1.3 Proposed method

However, except for Kaiser's eigenvalue greater than one rule, many methods mentioned above are not applicable to high dimensional data when the number of observed variables exceed the number of subjects, such that $p > n$, p is the number of observed variables (dimensions) and n is the number of subjects (sample size). Therefore, it boils down to the problem of how to determine the optimal projected dimensions, in other words, the number of latent factors, in a high dimensional setting. To this end, we propose a method of selecting optimal number of latent factors by utilizing Marchenko-Pastur law. Basically, this law originates from random matrix theory, which states that for a $n \times p$ random matrix following Wishart distribution,the limiting behaviour of its eigenvalues is generally bound by a constant ration $n/p \to c \in (0, \infty)$ as p approaches to ∞. Besides Kaiser's rule, Eigenvalue ratio method (Fan et al., 2021), Growth ratio method (Ahn and Horenstein, 2013) and Empirical Kaiser Criterion (Braeken and Van Assen, 2017) are chosen to compare with Marchenko-Pastur Law. The three methods are well specified to determine the optimal number of latent factors in high-dimensional data sets.

1.4 Overview

This paper will be arranged as follows, in next section, we introduce the basic structure of common factor model. A section of methods explanations is followed, in which Marchenko-Pastur Law, Eigenvalue ratio method, Growth ratio method and Empirical Kaiser Criterion are illustrated in detail. Then we perform simulation studies and two real data illustrations to further test the performance of MP law comparing with other four methods in high-dimensional data sets.

2 The Common Factor Model

The factor analytic model, in other words, the linear common factor model can be generally seen as a sparse modelling technique for covariance or correlation matrices(Peeters et al., 2012). The structure of factor analysis model is as follows: let $Y_1, \ldots, Y_n$ be n subjects with p observed variables for each Y_i, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $\xi_1, \ldots, \xi_n$ be n subjects with m realizations for each ξ_i, where ξ_i stands for latent common factors which have m dimensions (Peeters et al., 2012). In vector notations, $\boldsymbol{Y_i}^T \equiv [Y_{i1}, \ldots, Y_{ip}] \in \mathbb{R}^p$ represents the vector of a random subject and its realization can be denoted as $\mathbf{y_i}^T \equiv [y_{i1}, \ldots, y_{ip}] \in \mathbb{R}^p$. The realization of a random latent factor is denoted as $\boldsymbol{\xi_i}^T \equiv [\xi_{i1}, \ldots, \xi_{im}] \in \mathbb{R}^m$, which means there are m latent factors for a random $p-variate$ subject.

The factor analytic model for high dimensional data generally indicates a random p-*variate* subject Y_i is a linear combination of its m-*variate* latent factors ξ_i. Its mathematical form is as follows:

$$\underset{(p\times 1)}{\mathbf{y_i}} = \underset{(p\times m)}{\boldsymbol{\Lambda}} \cdot \underset{(m\times 1)}{\boldsymbol{\xi_i}} + \underset{(p\times 1)}{\boldsymbol{\epsilon_i}}, \tag{1}$$

where $p > m$, $\epsilon_i \in \mathbb{R}^p$ represents the error term, $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times m}$ stands for a matrix of factor loading with each element indicating the weights of jth variable associated with one unit change in l th latent factor, $l = 1, \ldots, m$. To guarantee the model structure, there are some assumptions to be maintained. Firstly, all subjects are independent, $y_i \perp y_{i'}$. Secondly, the factor loading matrix of $\boldsymbol{\Lambda}$ is of full rank of m. Thirdly, the latent factors and error term are normally distributed with mean of zero, where $\xi_i \sim \mathcal{N}(0, \boldsymbol{\Phi})$ and $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ as a diagonal matrix of positive values. Moreover, the correlation between latent common factors and error terms has to be zero for all i (Mulaik, 2009). The factor analytic model mainly focuses on covariance or correlation among observed variables, and the fundamental theorem of factor analysis can be worked out through some linear algebra operations:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}[\mathbf{y_i}\mathbf{y_i}^T] \\
&= \mathbb{E}[(\boldsymbol{\Lambda}\boldsymbol{\xi_i} + \boldsymbol{\epsilon_i})(\boldsymbol{\Lambda}\boldsymbol{\xi_i} + \boldsymbol{\epsilon_i})^T] \\
&= \boldsymbol{\Lambda}\boldsymbol{\xi_i}\boldsymbol{\xi_i}^T\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \\
&= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}
\end{aligned} \tag{2}$$

$\boldsymbol{\Sigma}$ denotes the covariance matrix of all the observed variables and this equation (2) demonstrates the fundamental theorem of factor analysis (Mulaik, 2009). Rearranging equation (2), $\boldsymbol{\Sigma} - \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T$ has the property of all off-diagonal elements are only related to the common latent factors. The principal diagonal part consists of the communalities of the observed variables, which are the variances of the observed variables related just to the common latent factors (Mulaik, 2009). Conventionally, the latent factors are assumed to be orthogonal. In this sense, for any non-singular matrix $\mathbf{H} \in \mathbb{R}^{m\times m}$, the fundamental theorem

of factor analysis in equation (2) can be expanded as

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}^T + \mathbf{\Psi} = (\mathbf{\Lambda H}) \left[\mathbf{H}^{-1}\mathbf{\Phi} \left(\mathbf{H}^T\right)^{-1}\right] (\mathbf{\Lambda H})^T + \mathbf{\Psi} \tag{3}$$

and this expanded equation (3) implies that there is an infinite number of matrices $\mathbf{\Lambda H}$ and $\mathbf{H}^{-1}\mathbf{\Phi} \left(\mathbf{H}^T\right)^{-1}$ can lead to the original covariance matrix $\mathbf{\Sigma}$.

3 Methods

3.1 Correlation matrix

Assume $\mathbf{y_i^T} \equiv [y_{i1}, \ldots, y_{ip}] \in \mathbb{R}^p$ is a realization of 1 observation and p variables from a random vector $\boldsymbol{Y_i}^T \equiv [Y_{i1}, \ldots, Y_{ip}] \in \mathbb{R}^p$. The sample covariance matrix can then be written as

$$\mathbf{S} = \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})^T$$

where $\bar{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} \mathbf{y_i}$. Then the sample correlation matrix can be easily obtained through $\mathbf{S}$ as

$$\mathbf{R} = \left[(\mathbf{S} \circ \mathbf{I})^{-\frac{1}{2}}\right] \mathbf{S} \left[(\mathbf{S} \circ \mathbf{I})^{-\frac{1}{2}}\right] \tag{4}$$

which is also regarded as the standardized sample covariance matrix. If original data $\mathbf{y_i} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, then standardized data follows $\mathcal{N}(0, \mathbf{\Sigma_R})$ where $\mathbf{\Sigma_R}$ is the population correlation matrix (Peeters et al., 2019). Generally, it is known that if $\mathbf{y_i}$ is multivariate normally distributed with $\mathcal{N}(\mu, \mathbf{\Sigma})$, its covariance matrix $\mathbf{S}$ then follows the Wishart distribution stated as $\mathbf{S} \sim \mathcal{W}_p(\frac{1}{n-1}\mathbf{\Sigma}, n-1)$, $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ denotes the scale matrix and $n-1$ is the degrees of freedom. Furthermore, since the correlation matrix $\mathbf{R}$ is standardized from $\mathbf{\Sigma}$ and is also multivariate normally distributed, we can safely assume that $\mathbf{R}$ follows Wishart distribution as well, $\mathbf{R} \sim \mathcal{W}_p(\frac{1}{n-1}\mathbf{\Sigma_R}, n-1)$, with $\mathbf{\Sigma_R}$ as the scale matrix and $n-1$ degrees of freedom.

3.2 Marchenko-Pastur Law

In this section, we introduce Marchenko-Pastur law, which is our proposed method that will be employed in selecting optimal dimensions in exploratory factor analysis (EFA).

Let $\mathbf{Y}$ be a $n \times p$ random matrix, n denotes the number of subjects and p denotes the number of observed variables. Each entry of $\mathbf{Y}$ is independently randomly drawn from multivariate normal distribution with mean of zero and variance of one, where $E(\mathbf{Y}) = 0$ and $E(\mathbf{YY^T}) = 1$. Hence, the sample covariance matrix of $\mathbf{Y}$ denotes $\mathbf{W}$ and

$$\mathbf{W} = \frac{1}{n}\mathbf{YY^T}$$

$\mathbf{W}$ has dimension of $p \times p$ and follows Wishart distribution. Denoting $\theta_1, \ldots, \theta_p$ as the eigenvalues of symmetric matrix $\mathbf{W}$, the empirical spectral distribution of $\mathbf{W}$ is then

$$F^W(y) := \frac{1}{p}\#\{1 \leq j \leq p : \theta_j(\mathbf{W}) < y\}$$

The limiting behaviour of all eigenvalues $\theta_j(\mathbf{W})$ extracted from sample covariance matrix $\mathbf{W}$ is determined by the Marchenko-Pastur law (Marčenko and Pastur, 1967). The average probability of eigenvalues is valid when the column dimension of $\mathbf{Y}$ approaches to infinity and its respect ratio with the number of subjects converges to a constant, such that $p \to \infty$ and $n/p \to c \in (0, \infty)$. To this end, the Marchenko-Pastur law states that the limiting spectral distribution of $\theta_j(\mathbf{W})$ is generalized as (Livan et al., 2018; Marčenko and Pastur, 1967)

$$f_c(y) = \frac{1}{2\pi y}\sqrt{(y - c_-)(c_+ - y)} \tag{5}$$

with $y \in (c_-, c_+)$ and the edge points $c_\pm$ are derived as $c_- = \left(1 - \sqrt{\frac{1}{c}}\right)^2$ and $c_+ = \left(1 + \sqrt{\frac{1}{c}}\right)^2$ and $c = n/p$. The derivation of Marchenko-Pastur scaling function shares the similar method as Wigner's semicircle law (Livan et al., 2018).

As the sample correlation matrix $\mathbf{R}$ is a sufficient representation of its population correlation matrix, and the fundamental theorem of factor model gives

$$\mathbf{\Sigma} - \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\mathbf{T} \tag{6}$$

$\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\mathbf{T}$ in equation (6) can also be regarded as reduced correlation matrix and it is assumed to be a Gramian and of rank m (Mulaik, 2009) , which is diagonalizable and its eigenvalues are non-negative (Peeters et al., 2019). Thus the selection criteria only depends on $c_+ = \left(1 + \sqrt{\frac{1}{c}}\right)^2$ and any extracted eigenvalues from sample correlation matrices smaller than c_+ are discarded.

3.3 Eigenvalue Ratio

Eigenvalue ratio is another method can be applied to determine the number of latent factors in high dimensional data. Specifically, the extracted eigenvalues from a sample correlation matrix $\mathbf{R}$ are arranged in descending order such that $\theta_1 \geqslant \theta_2 \geqslant \cdots \theta_{n \wedge p}$, where $\theta_{n \wedge p}$ denotes the last eigenvalue with index of $min\{n, p\}$ (Fan et al., 2021). The method can be demonstrated as follows, the ratio of two adjacent eigenvalues are calculated for all eigenvalues extracted from a sample correlation matrix, k_{max} denotes a predetermined random value for the number of latent factors in the data set, for instance, just as the same as in the case of confirmatory factor analysis, the number of latent factors is predetermined. Then the

optimal number of latent factors can be determined as

$$K_{ER} = \underset{i \leqslant k_{max}}{argmax} \frac{\theta_i}{\theta_{i+1}} \tag{7}$$

K_{ER} is the number of latent factors selected, representing the index of the largest eigenvalue ratio and the cut-off value is the maximum ratio of $\frac{\theta_i}{\theta_{i+1}}$ under the condition of $i \leqslant k_{max}$. (Ahn and Horenstein, 2013; Lam and Yao, 2012).

In the case of exploratory factor analysis, the number of latent factors in not known beforehand, which means k_{max} is unknown in terms of eigenvalue ratio method. An applicable estimation of the value for k_{max} could be the total number of eigenvalues extracted from sample correlation matrices, as it states that the signal eigenvalues can be decisively separately from the rest of eigenvalues when the eigenvalue ratio is the maximum of all ratios at $K_{ER} = k_i$, k_i is the index for the largest eigenvalue ratio (Fan et al., 2021).

3.4 Growth Ratio

Similar to eigenvalue ratio method, growth ratio method also employs the ratio from two adjacent eigenvalues, but in a different rationale. Particularly, Growth ratio method takes the regression residuals into account. In other words, growth ratio method considers eigenvalues that are not involved in the regression process.

The steps of growth ratio method begins with arranging eigenvalues extracted from sample correlation matrix $\mathbf{R}$ in descending order such that $\theta_1 \geqslant \theta_2 \geqslant \cdots \theta_{n \wedge p}$, where $\theta_{n \wedge p}$ denotes the last eigenvalue with index of $min\{n, p\}$, the growth ratio method is stated as

$$G_{GR} = \underset{i \leqslant g_{max}}{argmax} \frac{\ln\left[V(i-1)/V(i)\right]}{\ln\left[V(i)/V(i+1)\right]} = \frac{\ln\left(1+\theta_i^*\right)}{\ln\left(1+\theta_{i+1}^*\right)} \tag{8}$$

G_{GR} is the number of selected factors and the cut-off value is the maximum of $\frac{\ln(1+\theta_i^*)}{\ln(1+\theta_{i+1}^*)}$ under the condition of $i \leqslant g_{max}$ (Ahn and Horenstein, 2013). $V(i)$ is the possible sum of squared residuals from the regressions of response variables on the first i latent factors, which can be denoted as $V(i) = \sum_{g=i+1}^{g_{max}} \theta_g$. And θ_i^* can be written as $\theta_i^* = \theta_i / V(i)$ (Ahn and Horenstein, 2013; Onatski, 2010). g_{max} is the minimum value of n or p. The numerator and denominator of the growth ratio equation 8 both represent the growth rates of squared residuals from accounting for one less latent factors in the regressions (Ahn and Horenstein, 2013). The rationale behind growth ratio method is similar to eigenvalue ratio method, which means that the largest growth ratio can well separate the signal eigenvalues from the rest of noisy ones. Hence the index of the largest growth ratio can be regarded as the maximum number of latent factors to retain (Ahn and Horenstein, 2013; Onatski, 2010).

3.5 Empirical Kaiser Criterion

On the basis of Kaiser's rule (Kaiser, 1960) and Marchenko-Pastur law (Marčenko and Pastur, 1967), Empirical Kaiser Criterion (EKC) is proposed by Braeken and Van Assen (Braeken and Van Assen, 2017). One of the advantages of EKC is taking into account of the serial nature of eigenvalues from sample correlation matrices (Braeken and Van Assen, 2017). Basically, Empirical Kaiser Criterion develops an adaptive sequence of reference eigenvalues $\theta^{EKC} = \{\theta_1^{EKC}, \cdots, \theta_j^{EKC}\}$ (Braeken and Van Assen, 2017).

Empirical Kaiser Criterion consists three main ingredients. Firstly, its starting reference value follows Marchenko-Pastur law for which utilizing the positive edge point $c_+ = \left(1 + \sqrt{\frac{1}{c}}\right)^2$. Secondly, for subsequent eigenvalues, their reference values are calculated by adjusting the first reference value with proportional empirical correction, and the correction parameter is $\frac{p - \sum_{j=0}^{j-1} \theta_j}{(p-j+1)}$, p is the total number of observed variables and $\theta_0 = 0$. This correction parameter takes into account the serial nature of eigenvalues by taking the average remaining variance after accounting for the first up to the $(j-1)^{th}$ eigenvalues (Braeken and Van Assen, 2017). Moreover, the theoretical minimum value of θ_j should be 0. Thirdly, EKC requires that the eigenvalues should be greater than one. The theoretical rationale of this requirement originates from Kaiser (Kaiser, 1960), which argues that it is necessary and sufficient to have eigenvalues greater than one for a factor to have positive reliability (Kaiser, 1960). In combination of all the ingredients, the Empirical Kaiser Criterion is derived as

$$\theta_j^{EKC} = max \left[\frac{p - \sum_{j=0}^{j-1} \theta_j}{p - j + 1} \left(1 + \sqrt{\frac{1}{c}}\right)^2, \ 1 \right], with \ \ \theta_0 = 0 \tag{9}$$

The eigenvalue greater than its aspect reference value θ_j^{EKC} is retained as one of the common latent factors.

4 Simulation Study

In high-dimensional data settings, Kaiser's eigenvalue greater than one rule and eigenvalue ratio method can be applied to select the number of latent factors in factor analysis. Comparing with the Marchenko-Pastur law, the simulation study investigates the estimation accuracy for determining the number of latent factors in high-dimensional data of the five methods respectively.

The structure of simulation study design is aimed to include thoroughly possible combinations of the number of subjects n, the number of dimensions p, the number of latent factors m, latent factor loading λ and whether latent factors are orthogonal or not. This is a balanced design with each latent factor representing the same number of observed variables. In detail, we set the number of subjects to be $n = \{100, 500, 1000\}$ and the number of dimensions (observed variables) to be $p = \{200, 500, 1000\}$. The data is simulated in batches,

such that for each value of p, data sets with each value of m combined with each range of correlations ϕ of zero, low correlations of $(.10, .30)$ and high correlations of $(.75, .80)$ between latent factors that are also in conjunction with each of two scenarios of λ intervals, which are of low values in $(.60, .75)$ and high values in $(.75, .90)$ respectively. All the values from correlation intervals and factor loading intervals are generated uniformly, making them close to reality. Moreover, to get general estimating accuracy of each method, 100 random data sets are simulated from each combination of p, m, ϕ and λ with each number of subjects $n = \{100, 500, 1000\}$. Hence, with each value of p, 54 different combinations are generated. Correlation matrix is take from each simulated data set and subsequently all three methods are applied to each correlation matrix respectively. The estimating accuracy of each method is evaluated by the correct counts of each estimated number of latent factors. FMradio R package is utilized to simulate all the data.

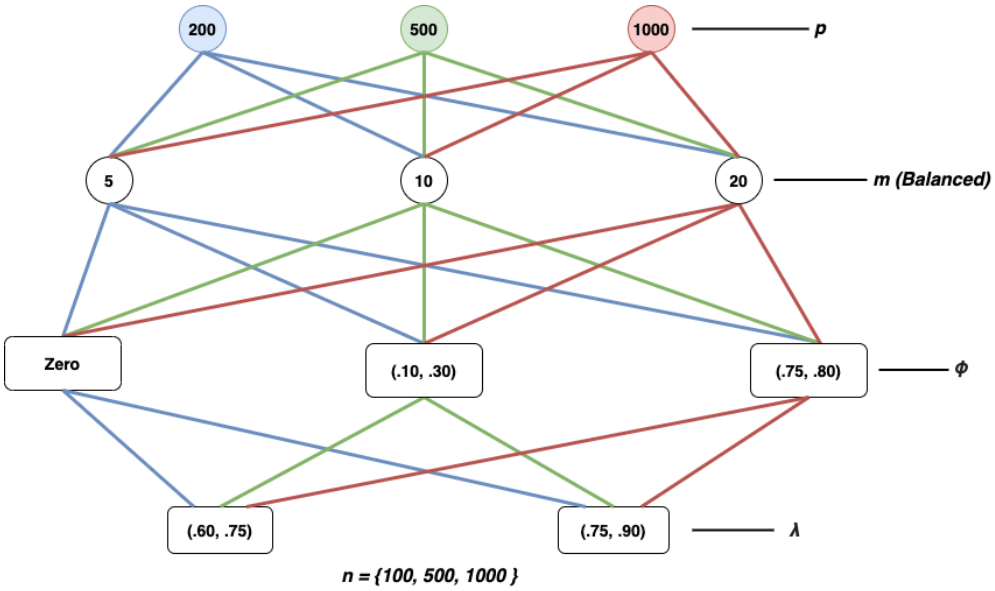


Figure 1: Simulation setup chart. A path connecting from top p to bottom λ indicates one combination of each value from p,m,ϕ and λ. 100 random data sets are simulated for each combination with each value of the number of subjects $n = \{100, 500, 1000\}$.Then all three methods are applied to each of the simulated correlation matrices.

5 Simulation Study Results

The results are displayed in three batches that are indexed by each value of $p = \{200, 500, 1000\}$. Figure 2, 3, 8, 9, 10, 11 show results for all combinations when $p = 200$, Figure 4, 5, 12, 13, 14, 15 display results when $p = 500$ and the last six figures (Figure 6, 7, 16, 17, 18, 19) show results when $p = 1000$. For all figures, left-hand side indicates the number of subjects n and the bottom represents the number of estimated latent factors m. On the right-hand side, five estimation methods are displayed. At the top header, "**m5**", "**m10**" and "**m20**" represent the predefined true values of m respectively.

Figure 2 and 3 with $Zero$ correlations between latent factors (Figures 8 to 11 with $High$ or

Low correlations between latent factors are in Appendix A) demonstrate the counted times of estimated values of m with Marchenko-Pastur law (MP), Eigenvalue ratio (EigRatio), Kaiser's rule (Kaiser), Empirical Kaiser(EmpKaiser) and Growth ratio method, comparing to the true values of $m = \{5, 10, 20\}$ when $p = 200$. Firstly, Empirical Kaiser's criterion and Marchenko-Pastur law perform equally well with fully correct estimates in all scenarios, except when the true $m = 20$ and the sample size $n = 100$, Empirical Kaiser criterion performs slightly better than MP law, providing more correct estimates of m equalling to 20. Moreover, Eigenvalue ratio and Growth ratio method have the same performance for all combinations with fully correct estimates when sample size $n = 500$ and $n = 1000$ respectively, and overestimation of the values of m when sample size $n = 100$. On the other hand, Kaiser's rule can estimate the values of m with full correctness when sample size n equals to either 500 or 1000 with high values of factor loadings λ. Other than that, Kaiser's rule tends to overestimate m values in most of the scenarios.

Figure 2: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio method. $p = 200$, ϕ is *Zero* and λ is of *Low* values from $(.60, .75)$

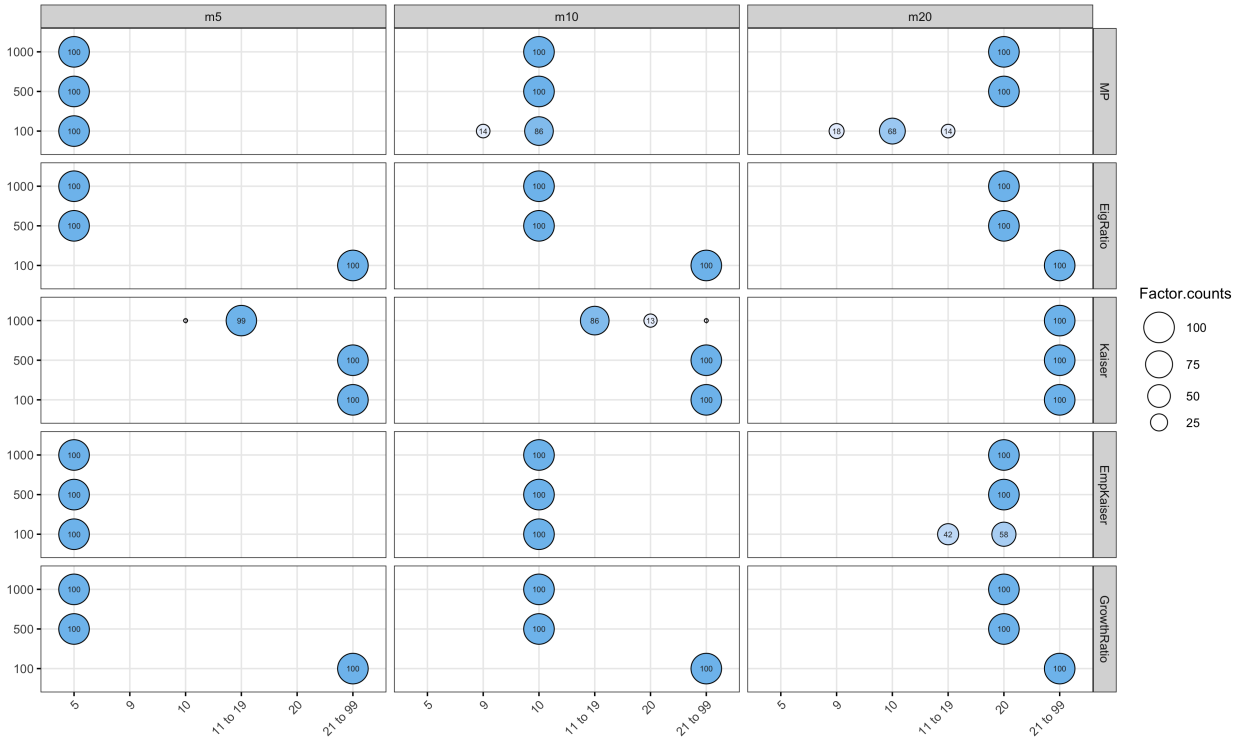


Figure 3: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 200$, ϕ is $Zero$ and λ is of $High$ values from $(.75, .90)$

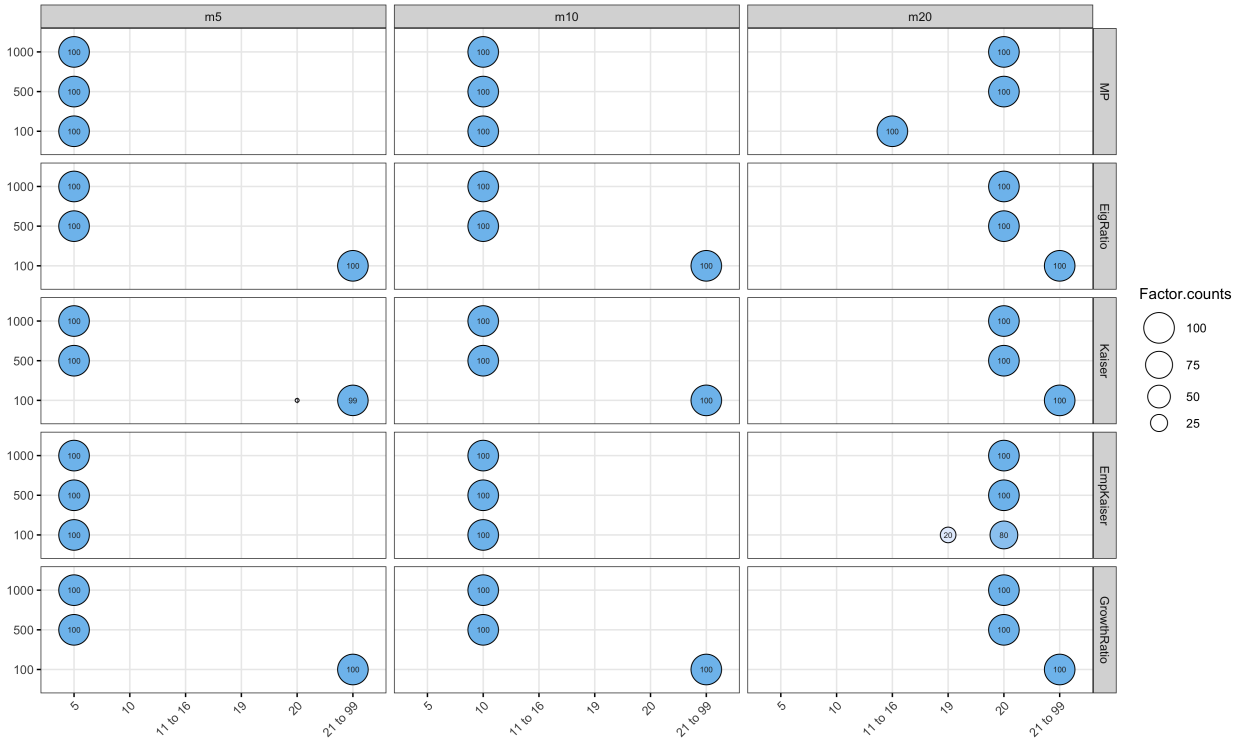


Figure 4 and 5 with $Zero$ correlations between latent factors (Figures 12 to 15 with $High$ or Low correlations between latent factors are in Appendix B) display the counted times of estimated m values by means of MP law, Eigenvalue Ratio, Kaiser's rule, Empirical Kaiser criterion and Growth ratio method when $p = 500$, comparing with the predefined true values of $m = \{5, 10, 20\}$. When factor loadings λ are of high values, Empirical Kaiser criterion can estimate all m values fully correctly. MP law estimates all correctly except for when sample size $n = 100$ and true $m = 20$, where its estimated m values are not exactly the predefined value 20, but as close as possible in most time such as 17, 18 and 19. Growth ratio method can estimate m values all correctly when sample size $n = 500$ and $n = 1000$ respectively, but overestimate largely when $n = 100$. Eigenvalue ratio method can estimate all values of m fully correctly only when sample size $n = 1000$ and overestimate in other cases. As to Kaiser's rule, it largely overestimates the values of m when factor loadings are low. Even when factor loadings are high, it performs only slightly better when sample size $n = 1000$.

Figure 4: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is $Zero$ and λ is of Low values from $(.60, .75)$

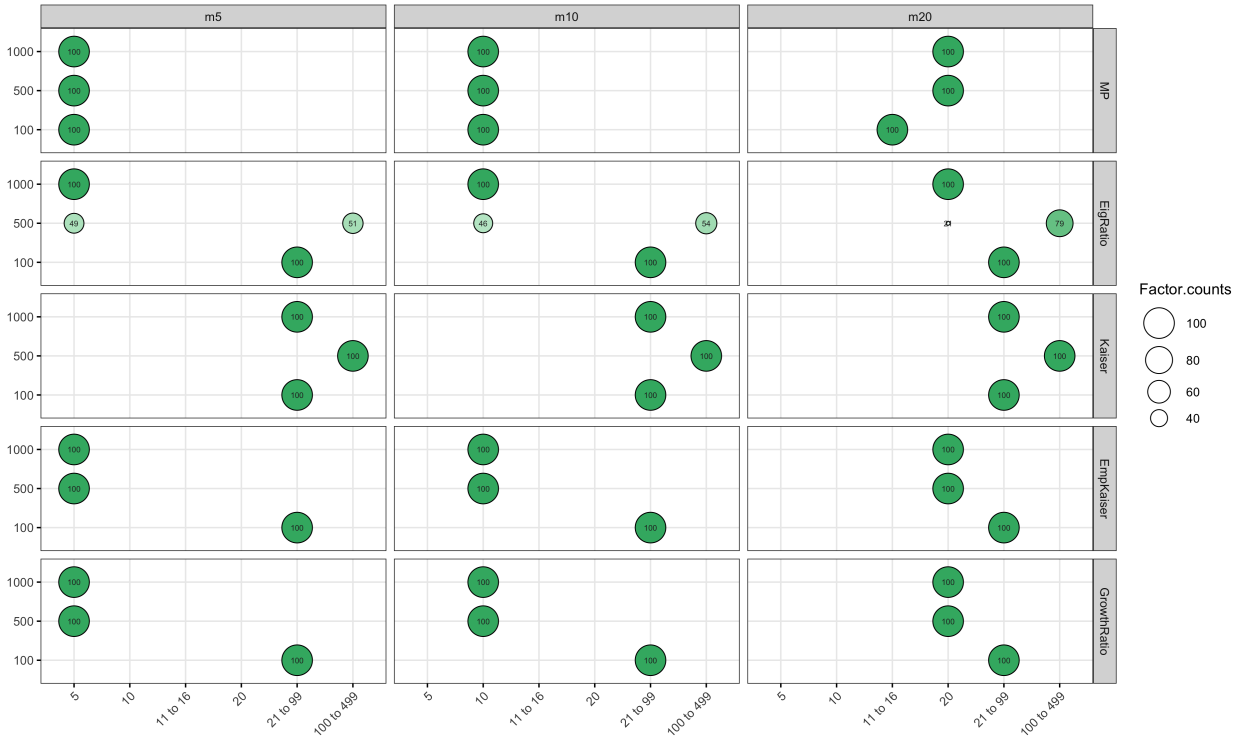


Figure 5: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is $Zero$ and λ is of $High$ values from $(.75, .90)$

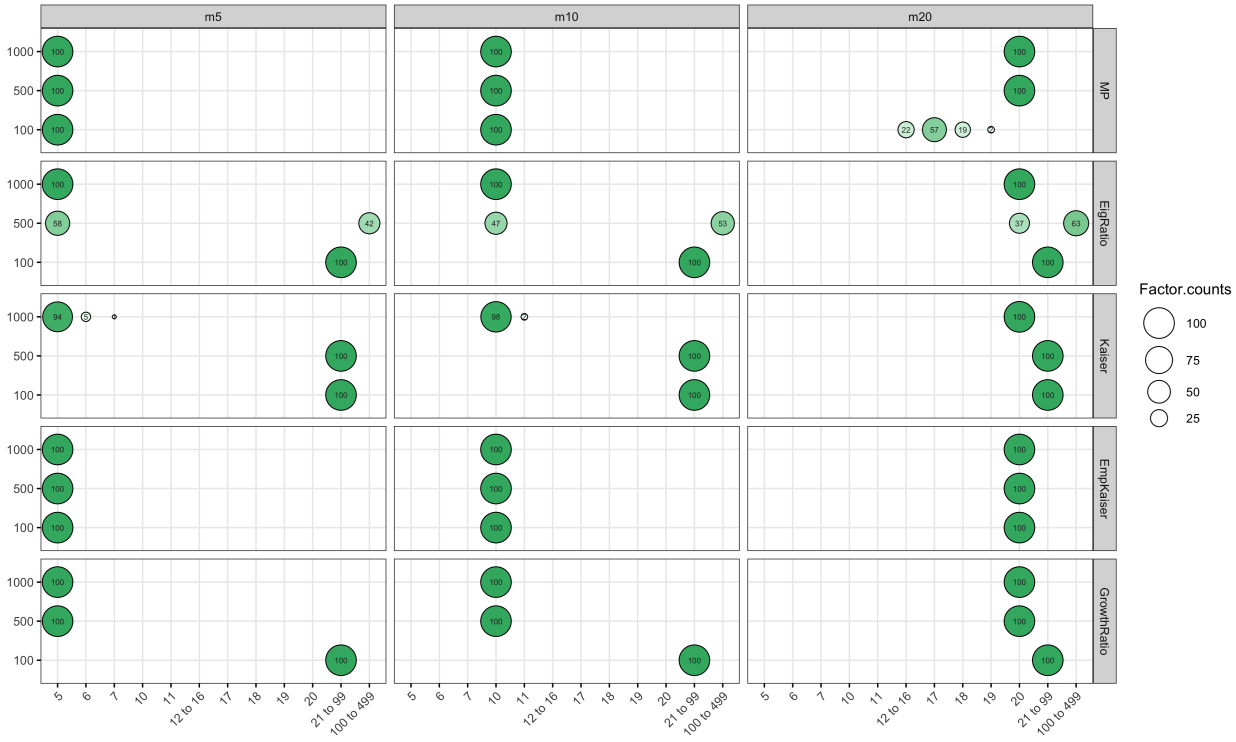


Figure 6 and 7 with *Zero* correlations between latent factors (Figures 16 to 19 with *High* or *Low* correlations between latent factors are in Appendix C) exhibit the counted times of estimated number of latent factors by means of MP law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio method when $p = 1000$, comparing with the predefined true values of m equalling to $\{5, 10, 20\}$. In this setting, Marchenko-Pastur law performs the best compared to the other methods, especially when sample size $n = 500$ and $n = 1000$ respectively. Though MP law cannot estimate m values fully correctly when $n = 100$ and predefined true $m = 20$, its estimated values for m are highly close to the true m value 20, such as 17,18 or 19. On the other hand, Empirical Kaiser criterion largely overestimates m values when sample size $n = 100$. Growth ratio method performs slightly better than Eigenvalue ratio method when sample size $n = 1000$, where Growth ratio method can estimate with 100 times correctness, but Eigenvalue ratio method can only estimate roughly half times correctly. When sample size $n = 100$ or $n = 500$, both methods tend to overestimate m values to a large extent. In terms of Kaiser's rule, it fails to provide the correct m values for all cases.

Figure 6: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is *Zero* and λ is of *Low* values from $(.60, .75)$

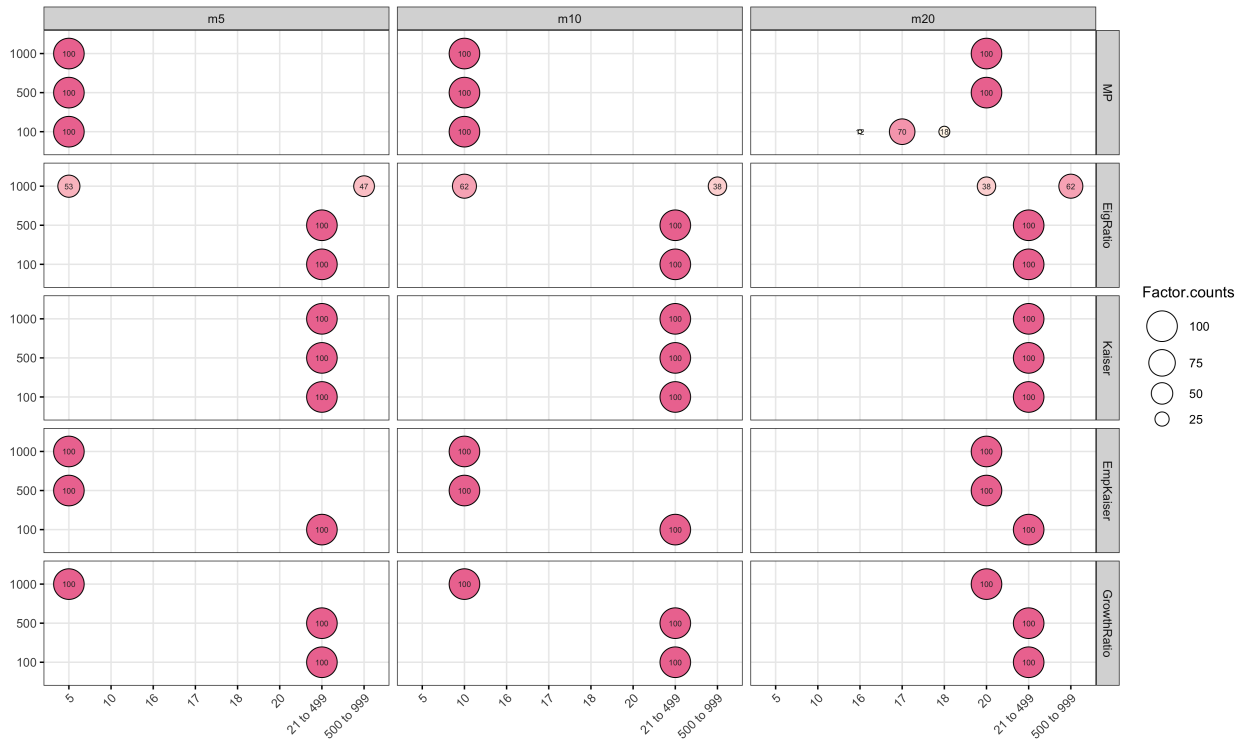
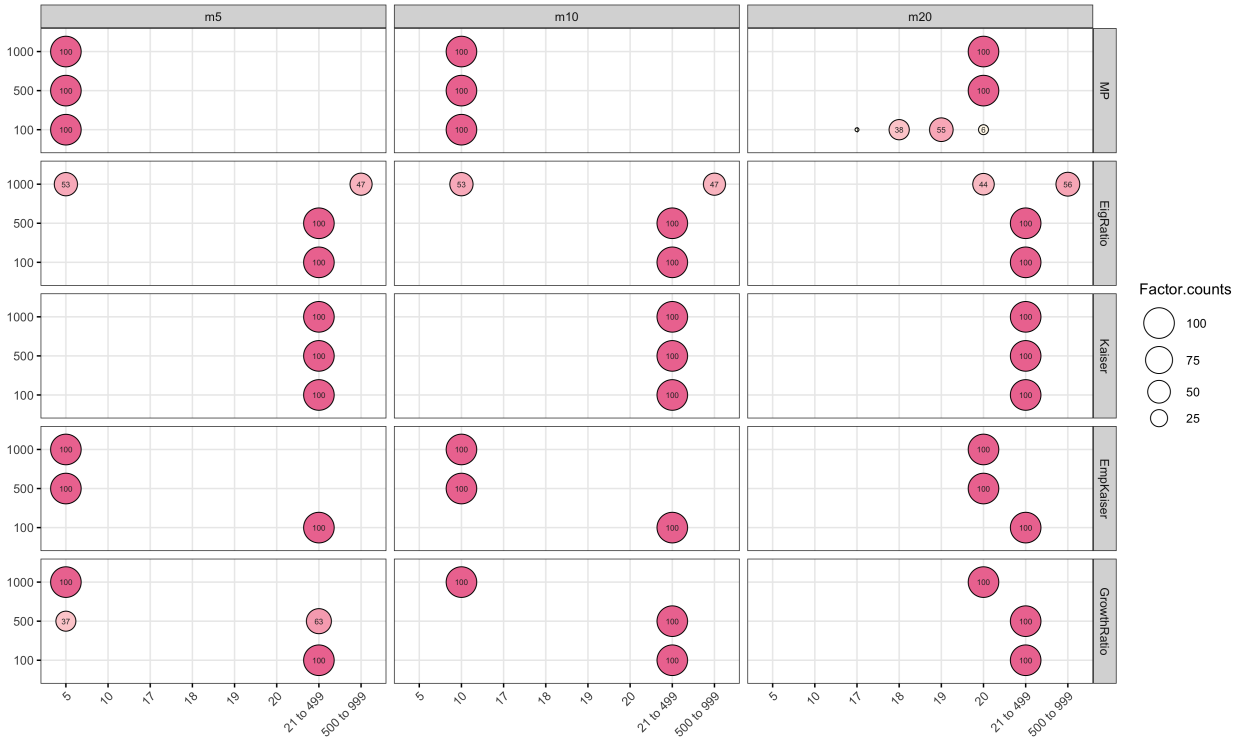


Figure 7: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is $Zero$ and λ is of $High$ values from $(.75, .90)$



6 Discussion

In general, Marchenko-Pastur law outperforms the other four methods in most of the cases. The only exception is when dimension $p = 500$. In this case, Empirical Kaiser Criterion can provide fully correct estimations for m values when factor loadings λ is set to be high. Although Marchenko-Pastur law does not estimate fully correctly in the same case when sample size $n = 100$, its estimated m values are close to the prespecified true m value equalling 20, such estimated m values are 17, 18 or 19. The performance of Marchenko-Pastur law is prominent especially when dimension is much larger than sample sizes. For instance, Empirical Kaiser Criterion completely fails to provide any correct estimations of m values in the scenario of $p = 1000$ and $n = 100$. This indicates that p/n ratio has an imperative influence on the performance of MP law. Moreover, the correlations ϕ between common latent factors have no obvious impacts on the performance of MP law method. In multiple previous studies, the independence assumption of entries from random matrix is weakened or relaxed. In the original paper by Marchenko and Pastur (Marčenko and Pastur, 1967), they relax the independent entries to independent rows. Götze and Tikhomirov study sample covariance matrices for which these matrices suffice the conditions of martingale-type and no assumptions are made on the independence of entries (Gotze and Tikhomirov, 2006; Götze and Tikhomirov, 2007). For our simulation study, MP law method demonstrates prominent estimation accuracy among all five methods, even on the conditions of high correlations(ϕ)

between latent factors.

Empirical Kaiser Criterion (EKC) is a competitive candidate to Marchenko-Pastur law when dimension p is moderate comparing to its respect samples size. Specifically, when dimension $p = 200$, Empirical Kaiser Criterion performs slightly better than MP law with a few more accurate estimates of m values when sample size $n = 100$. However, the performance of EKC becomes less accurate as dimension p increases. For instance, when dimension $p = 1000$ is much higher than sample size $n = 100$, EKC overestimates m values to a large extent, whereas MP law only slightly underestimates m values with close estimations, such as 17, 18 or 19. In addition, high values of factor loadings also assist EKC's performance, but in real data cases, values of factor loadings are seldom as high as our simulation setups. Furthermore, even though EKC performs better to a certain extend than MP law, the efficiency of EKC reduces to a large extent and its running time rises substantially when p increases.

As to eigenvalue ratio and growth ratio method, their performances on estimating the number of latent factors are relatively similar. Both p/n ratio and factor loadings λ play important roles. Basically, as long as the number of observed variables exceeds the sample size, $p > n$, both eigenvalue ratio and growth ratio method turn to overestimate m values. This problem stems from the choice of k_{max}, which is obviously as many as the smaller value of sample size n or dimension p, in which $min(n, p)$. The cut-off ratios in both methods depend sensitively on k_{max} (Ahn and Horenstein, 2013). A prespecified k_{max} might solve this problem, but contradicts to the case of exploratory factor analysis.

Kaiser's rule fails in most of the scenarios, particularly when observed dimensions exceeding sample sizes, $p > n$. Estimated latent dimensions from Kaiser's rule are generally not plausible in high-dimensional data sets.

7 Real Data Illustrations

In this section, two real high dimensional data sets are employed to check if Marchenko-Pastur Law performs the best among all the selected methods.

7.1 Small, Round Blue Cell Tumors Gene expression Data

Small, round blue cell tumors (SRBCTs), including neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), share the same name for the reason of their similar appearance on routine histology (Khan et al., 2001; Robison et al., 1997). Difficulties arise from distinguishing and precisely diagnosing cancers among these types. In practice, current diagnosis techniques, such as immunohistochemistry and molecular techniques, have drawbacks of either the examination of only one protein at a time or failure to provide precise diagnosis (Khan et al., 2001). Gene-expression data generated from cDNA microarrays can mitigate these problems by allowing analyzing

multiple signal markers at the same time. Subsequently, artificial neural networks can be utilized to categorize SRBCTs into different subgroups (Khan et al., 2001).

The SRBCTs gene-expression data is collected by cDNA microarrays, which are prepared, probe labelled, hybridized, and image acquired according to the standard NHGRI protocol. The cDNA clones are obtained from Research Genetics (Huntsville, Alabama). In total, cDNA microarrays contains 6567 genes, which consist of 3789 sequence-verified known genes and 2778 sequence-verified ESTs (Khan et al., 2001). The gene-expression data is further selected through filtering with red intensity value greater then 20 across all experiments, keeping 2308 genes in the data set. Training sample size is 63, which includes tumor biopsy material and cell lines (Khan et al., 2001). Relative red intensity is applied on all the selected gene-expression data by calculating RRI = mean intensity of that spot/mean intensity of filtered genes.

According to the original paper (Khan et al., 2001), a principle component analysis (PCA) is applied on the training data set to reduce the burden of high dimensionality. They find out that the first 10 principle components dominate the projected space for each sample in training data and account for approximately 63% of the total variance. Their justification for selecting 10 principle components considers that the remaining variance in the data matrix are unrelated to classifying the four cancers. For our analysis, our aim is to check if Marchenko-Pastur law can provide the most accurate latent dimensions comparing to the other four methods, EKC, eigenvalue ratio method, growth ratio method and Kaiser's rule. The same training data set of SRBCTs undergoes each of the factor selection methods, and the results are shown in Table 1, which displays that in the real high-dimensional data setting with sample size $n = 63$ and observed variables $p = 2308$, Marchenko-Pastur law generates 11 latent dimensions, on the other hand, empirical Kaiser criterion, eigenvalue ratio, growth ratio and Kaiser's rule all overestimate the number of latent dimensions to a large extend. With 11 selected latent factors, a factor analysis is performed on SRBCTs gene-expression data. R package FMradio is used and the method employs a Varimax-rotated maximum likelihood approach that is applied on the redundancy filtered correlation matrix of SRBCTs gene-expression data (Peeters et al., 2019). All 11 factors explain 65% of variance, which is almost the same as 63% of variance covered by the 10 retained PCA components in the paper of Khan et al., 2001. Table 2 displays the resulted cumulative variance and proportion explained from factor analysis on redundancy-filtered correlation matrix. In this case, Marchenko-Pastur law indeed performs the best in estimating latent dimensions. The benefits of applying MP law are its easiness, efficiency and accuracy, instead of constantly calibrating PAC components and tuning different parameters for ANN model.

	Methods of selecting latent dimensions				
	Marchenko-Pastur	EKC	Eigenvalue ratio	Growth ratio	Kaiser's rule
Estimated m	11	61	62	62	61

Table 1: Comparison of estimated latent dimensions by means of Marchenko-Pastur law, EKC, Eigenvalue ratio method, Growth ratio method and Kaiser's rule for SRBCTs gene-expression data with sample size $n = 63$ and observed dimensions $p = 2308$. Marchenko-Pastur law provides the most accurate estimate of latent dimensions compared to the retained PCA components in the paper of Khan et al., 2001

	Latent factors										
	1	2	3	4	5	6	7	8	9	10	11
Cumulative variance	0.04	0.07	0.09	0.11	0.13	0.15	0.16	0.17	0.18	0.19	0.20
Proportion explained	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01

Table 2: SRBCTs data: Cumulative variance and proportion explained for each latent factors when $m = 11$. The sequence of latent factors are ordered by its explained proportion of total variance. All 11 factors account for 20% of variance.

7.2 Head and Neck Squamous Cell Carcinoma Data

In this section, another real data set of head and neck squamous cell carcinoma (HNSCC) is employed to check if Marchenko-Pastur law can estimate the true number of latent factors. The data is available from 175 head and neck squamous cell carcinoma patients in Amsterdam University medical center. Features of shape, intensity, texture and wavelet-type are extracted by feature extraction methods with 3D implementation. For wavelet-type features, a wavelet transform is obtained through a Coiflet scaling function. The images of texture and wavelet are discretized to a fixed bin size of 64 bins and $0.25gml^{-1}$ standardized uptake value (SUV) (Peeters et al., 2019; van Velden et al., 2016). In total 432 rediomics features are extracted for each of the patient in the sample. Moreover, some features are removed from the data because they don't contribute variations to the respective patients (Peeters et al., 2019). Additionally, one patient is left out from the data set because of 0 follow-up time. Thus, the final data set has sample size $n = 174$ and observed features $p = 432$. According to the paper of Peeters et al., 2019, the number of latent factors is estimated to be 8 and the Guttman Bound suggests that the upper bound of optimal number of latent dimensions ought to be 13 (Peeters et al., 2019). In addition, the 8 latent factors account for approximately 76% of variance in the data set.

For our analysis of determining the optimal number of latent dimensions, the normalized

head and neck squamous cell carcinoma data goes through each of the latent dimension selection methods mentioned above. The results are displayed in Table 3. It indicates that Marchenko-Pastur law and Empirical Kaiser Criterion (EKC) both provide close values of latent dimensions to the original estimate of 8. However, EKC tends to overestimate the number of latent factors. A further factor analysis with 7 and 9 factors are applied respectively to the head and neck squamous cell carcinoma data. The similar method as above of Varimax-rotated maximum likelihood approach is performed and the results are displayed in Table 4 and Table 5 respectively. Results indicate that 7 factors explain 78% of variance in the data set (Table 4) and 9 factors account for 82% of variance (Table 5). The extra two factors only contribute slightly to explaining the total variance in the data set and their proportion explained stay the same.

	Methods of selecting latent dimensions				
	Marchenko-Pastur	EKC	Eigenvalue ratio	Growth ratio	Kaiser's rule
Estimated m	7	9	167	167	18

Table 3: Comparison of estimated latent dimensions by means of Marchenko-Pastur law, EKC, Eigenvalue ratio method, Growth ratio method and Kaiser's rule for head and neck squamous cell carcinoma data with sample size $n = 174$ and observed features $p = 432$. Marchenko-Pastur law provides the most accurate estimate of latent dimensions compared to the optimal dimensions of latent vectors in the paper of Peeters et al., 2019

	Latent factors						
	1	2	3	4	5	6	7
Cumulative variance	0.22	0.43	0.56	0.66	0.70	0.74	0.78
Proportion explained	0.22	0.22	0.12	0.10	0.04	0.04	0.04

Table 4: HNSCC data: Cumulative variance and proportion explained for each latent factors when $m = 7$. The sequence of latent factors are ordered by its explained proportion of total variance. All 7 factors account for 78% of variance.

	Latent factors								
	1	2	3	4	5	6	7	8	9
Cumulative variance	0.24	0.44	0.57	0.64	0.69	0.73	0.76	0.80	0.82
Proportion explained	0.24	0.20	0.13	0.07	0.05	0.04	0.04	0.03	0.02

Table 5: HNSCC data: Cumulative variance and proportion explained for each latent factors when $m = 9$. The sequence of latent factors are ordered by its explained proportion of total variance. All 9 factors account for 82% of variance.

In general, we see the distinct advantages of Marchenko-Pastur law, especially when the number of observed dimensions exceeds sample sizes. Moreover, correlations between latent factors and values of factor laodings have little influence on the estimation accuracy of Marchenko-Pastur law, which also imply its benefits of easiness to apply and high estimation accuracy, comparing to other dimension selection methods in explanatory factor analysis.

Our research sets some new areas to further explore and improve. In our study, we limit our research only on raw correlation matrices, in other words, the correlation matrices extracted from raw simulated data sets. For further researches, ill-behaved or singular raw correlation matrices could be regularized by penalized maximum log likelihood approach. The problem then boils down to if Marchenko-Pastur law performs well on regularized correlation matrices. Another research direction could be what are the bounding values that the average eigenvalues probability distribution converge to. Lastly, the simulation study could be set to unbalanced structure, where each latent factor covers different number of observed variables.

References

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.

Braeken, J., & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods*, *22*(3), 450.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245–276.

Child, D. (2006). *The essentials of factor analysis*. A&C Black.

Fan, J., Wang, K., Zhong, Y., & Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *36*(2), 303.

Gotze, F., & Tikhomirov, A. (2006). Limit theorems for spectra of positive random matrices under dependence. *Journal of Mathematical Sciences*, *133*(3), 1257–1276.

Götze, F., & Tikhomirov, A. N. (2007). Limit theorems for spectra of random matrices with martingale structure. *Theory of Probability & Its Applications*, *51*(1), 42–64.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate behavioral research*, *17*(2), 193–219.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological measurement*, *66*(3), 393–416.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141–151.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679.

Kim, J.-O., Ahtola, O., Spector, P. E., Mueller, C. W., et al. (1978). *Introduction to factor analysis: What it is and how to do it*. Sage.

Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 694–726.

Livan, G., Novaes, M., & Vivo, P. (2018). Introduction to random matrices theory and practice. *Monograph Award*, 63.

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457.

Mulaik, S. A. (2009). *Foundations of factor analysis*. CRC press.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, *92*(4), 1004–1016.

Peeters, C. F. et al. (2012). *Bayesian exploratory and confirmatory factor analysis: Perspectives on constrained-model selection*. Utrecht University.

Peeters, C. F., Übelhör, C., Mes, S. W., Martens, R., Koopman, T., de Graaf, P., van Velden, F. H., Boellaard, R., Castelijns, J. A., Beest, D. E. t., et al. (2019). Stable prediction with radiomics data. *arXiv preprint arXiv:1903.11696*.

Pett, M. A., Lackey, N. R., Sullivan, J. J., et al. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. sage.

Robison, L., Pizzo, P., & Poplack, D. (1997). Principles and practice of pediatric oncology.

van Velden, F. H., Kramer, G. M., Frings, V., Nissen, I. A., Mulder, E. R., de Langen, A. J., Hoekstra, O. S., Smit, E. F., & Boellaard, R. (2016). Repeatability of radiomic features in non-small-cell lung cancer [18f] fdg-pet/ct studies: Impact of reconstruction and delineation. *Molecular imaging and biology*, *18*(5), 788–795.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. *Problems and solutions in human assessment*, 41–71.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, *99*(3), 432.

Appendix A Counts of estimated m values when p = 200

Figure 8: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 200$, ϕ is of *Low* values from $(.10, .30)$ and λ is of *Low* values from $(.60, .75)$

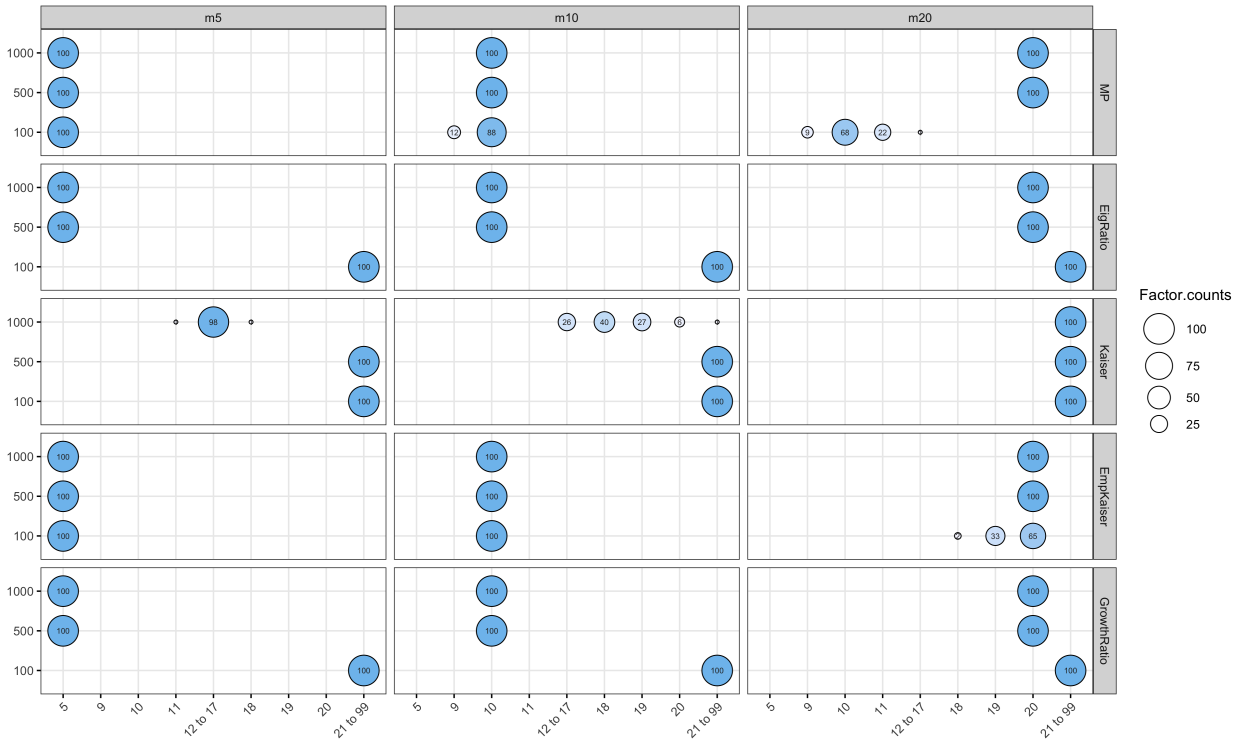


Figure 9: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 200$, ϕ is of Low values from $(.10, .30)$ and λ is of $High$ values from $(.75, .90)$

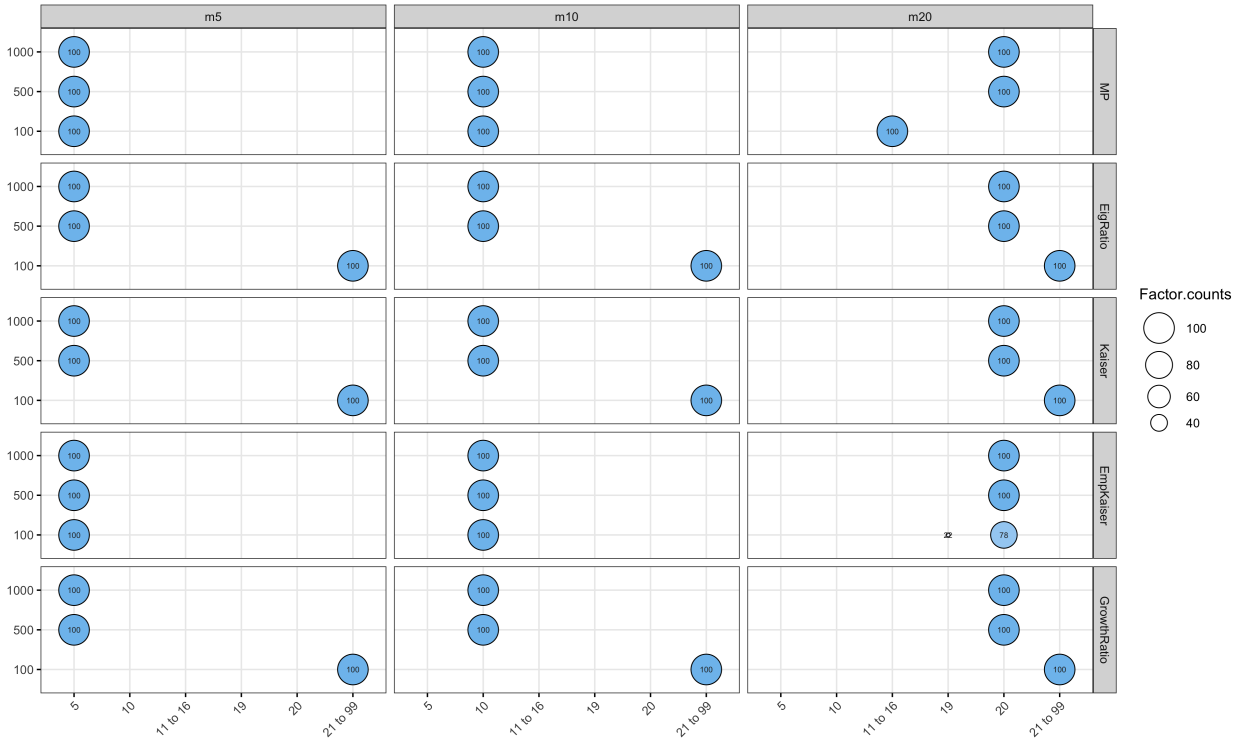


Figure 10: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 200$, ϕ is of $High$ values from $(.75, .80)$ and λ is of Low values from $(.60, .75)$

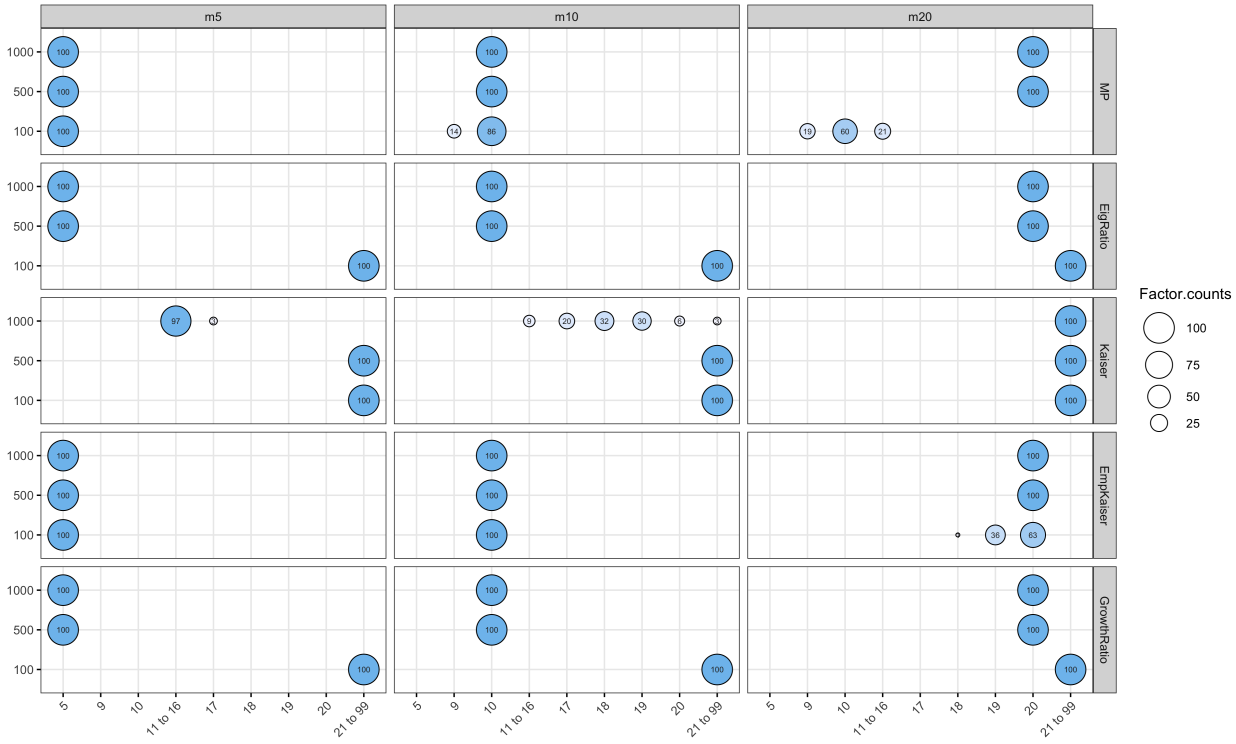
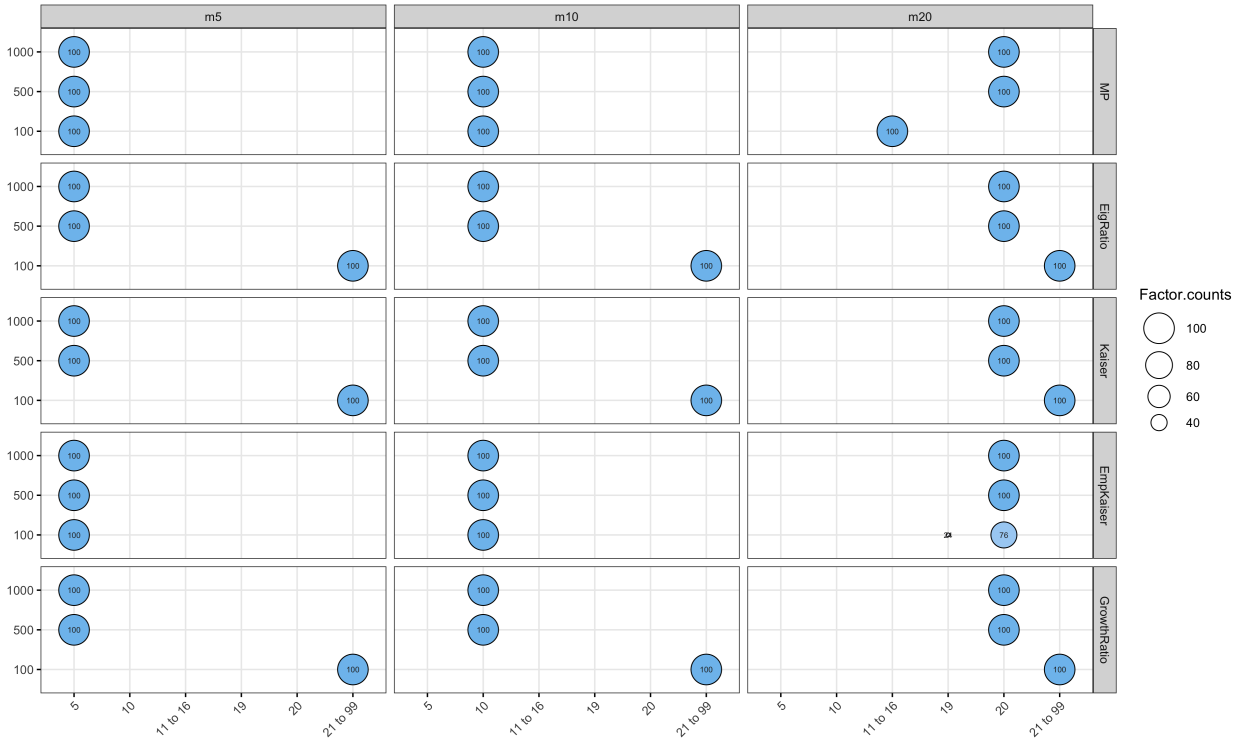


Figure 11: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 200$, ϕ is of $High$ values from $(.75, .80)$ and λ is of $High$ values from $(.75, .90)$



Appendix B Counts of estimated m values when p = 500

Figure 12: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is of *Low* values from $(.10, .30)$ and λ is of *Low* values from $(.60, .75)$

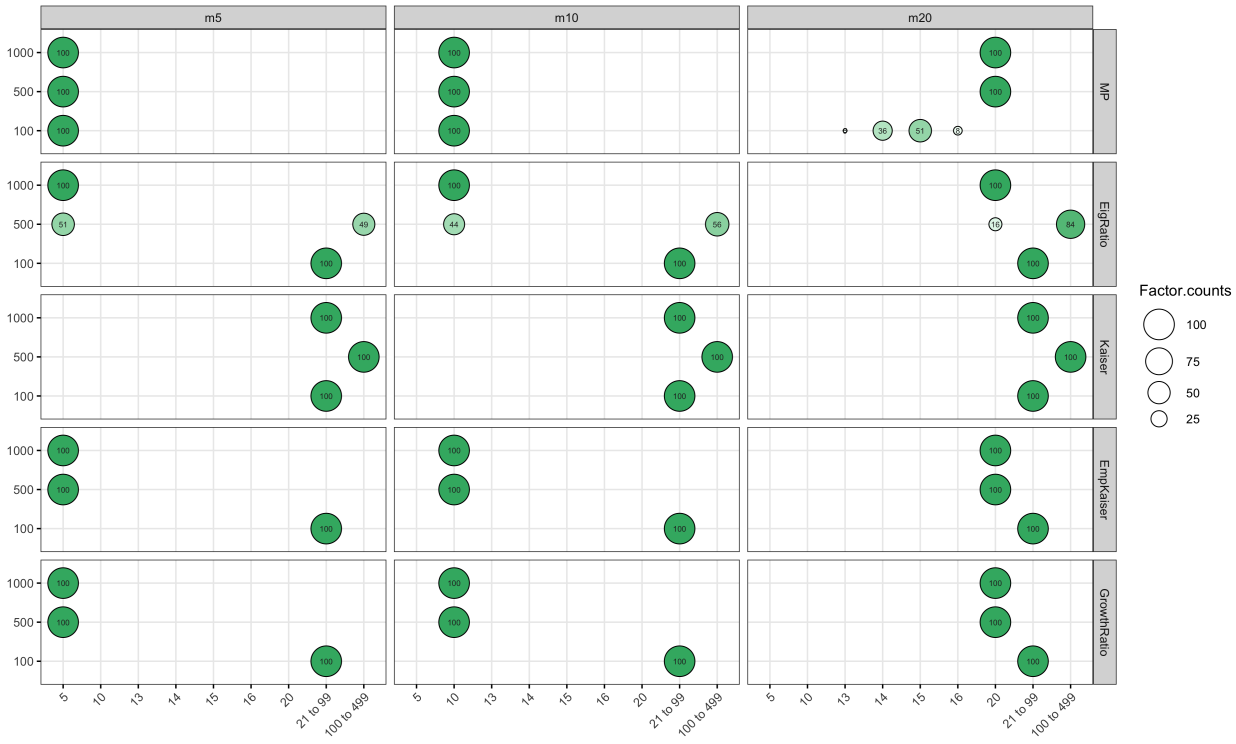


Figure 13: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is of Low values from $(.10, .30)$ and λ is of $High$ values from $(.75, .90)$

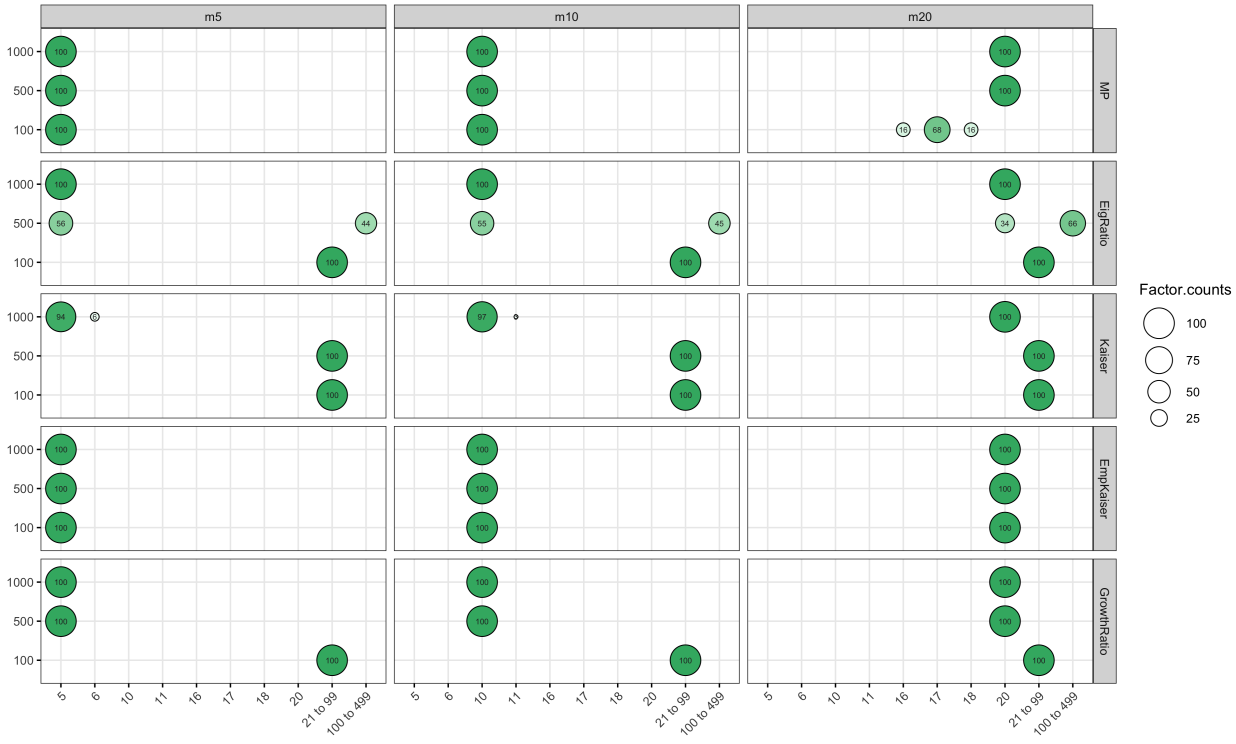


Figure 14: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is of $High$ values from $(.75, .80)$ and λ is of Low values from $(.60, .75)$

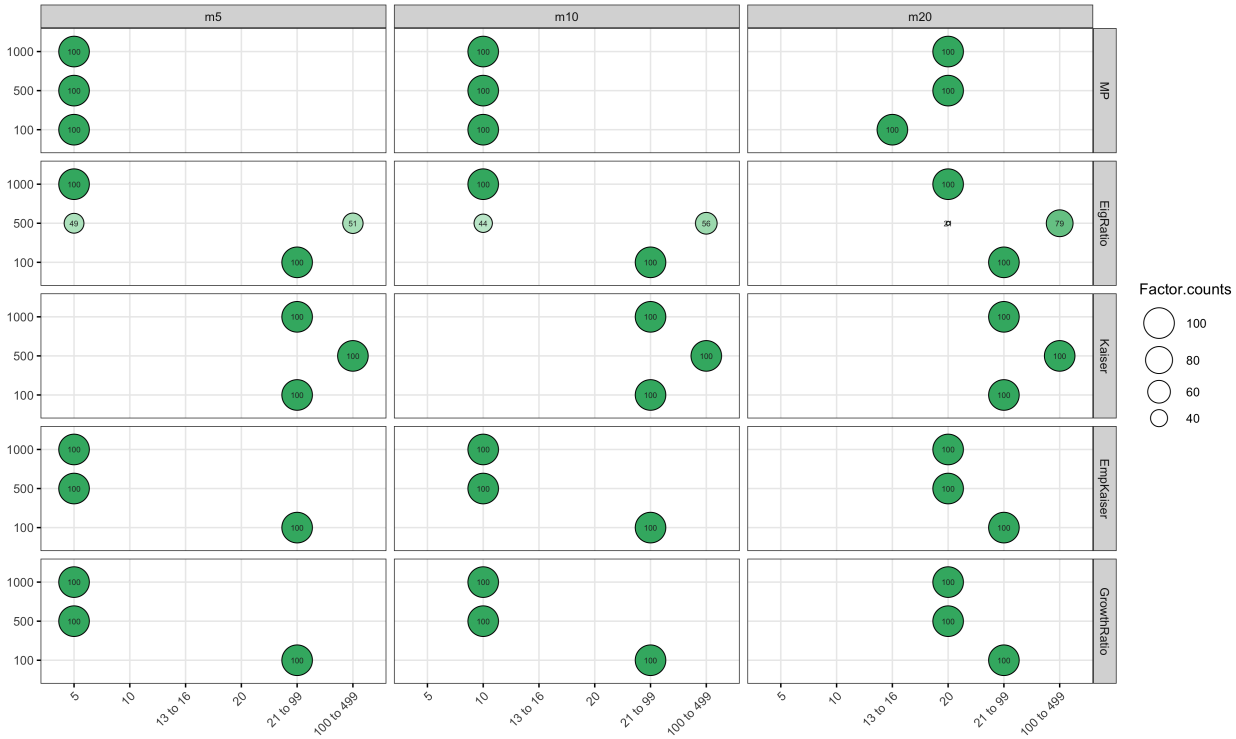
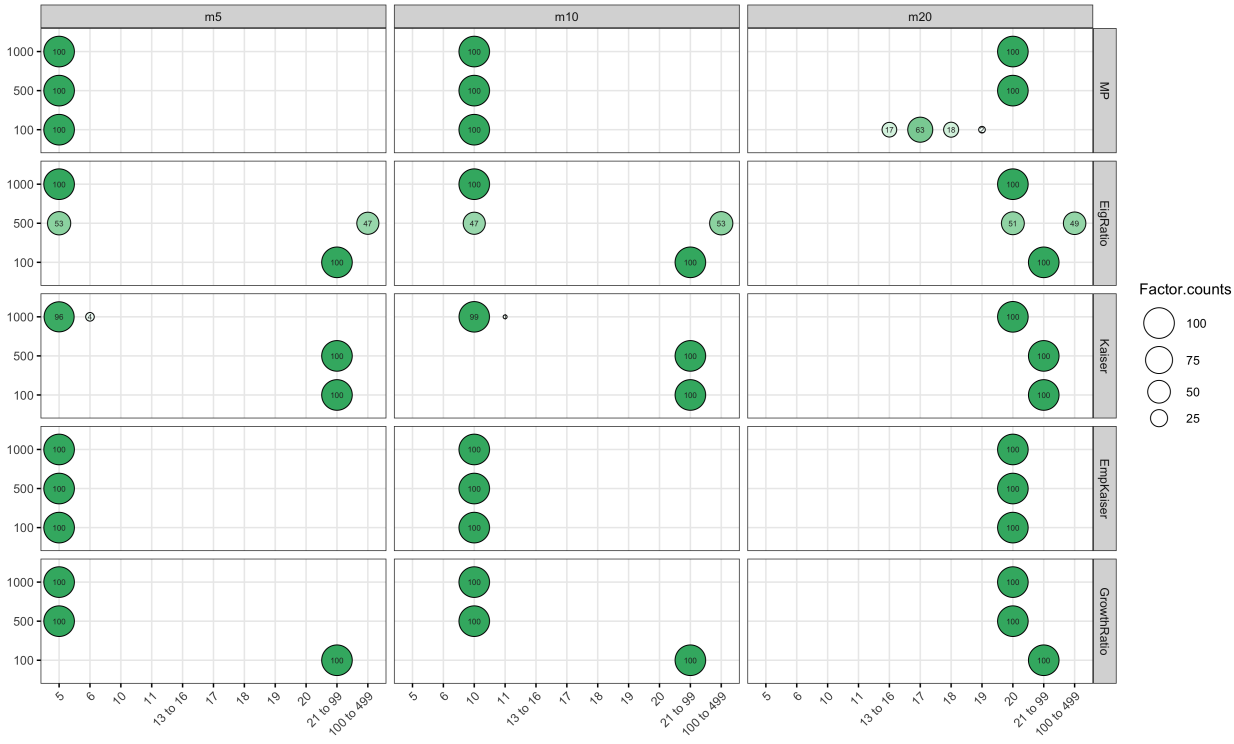


Figure 15: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 500$, ϕ is of $High$ values from $(.75, .80)$ and λ is of $High$ values from $(.75, .90)$



Appendix C Counts of estimated m values when p = 1000

Figure 16: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is of *Low* values from $(.10, .30)$ and λ is of *Low* values from $(.60, .75)$

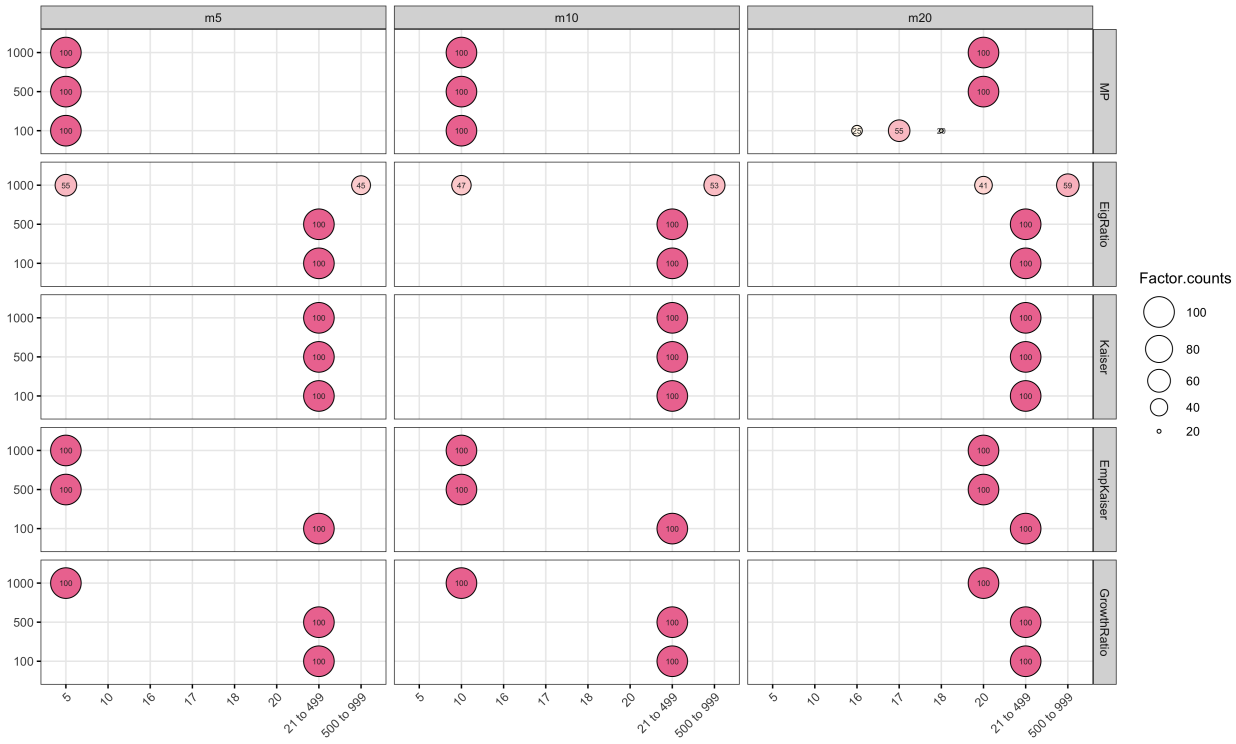


Figure 17: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is of Low values from $(.10, .30)$ and λ is of $High$ values from $(.75, .90)$

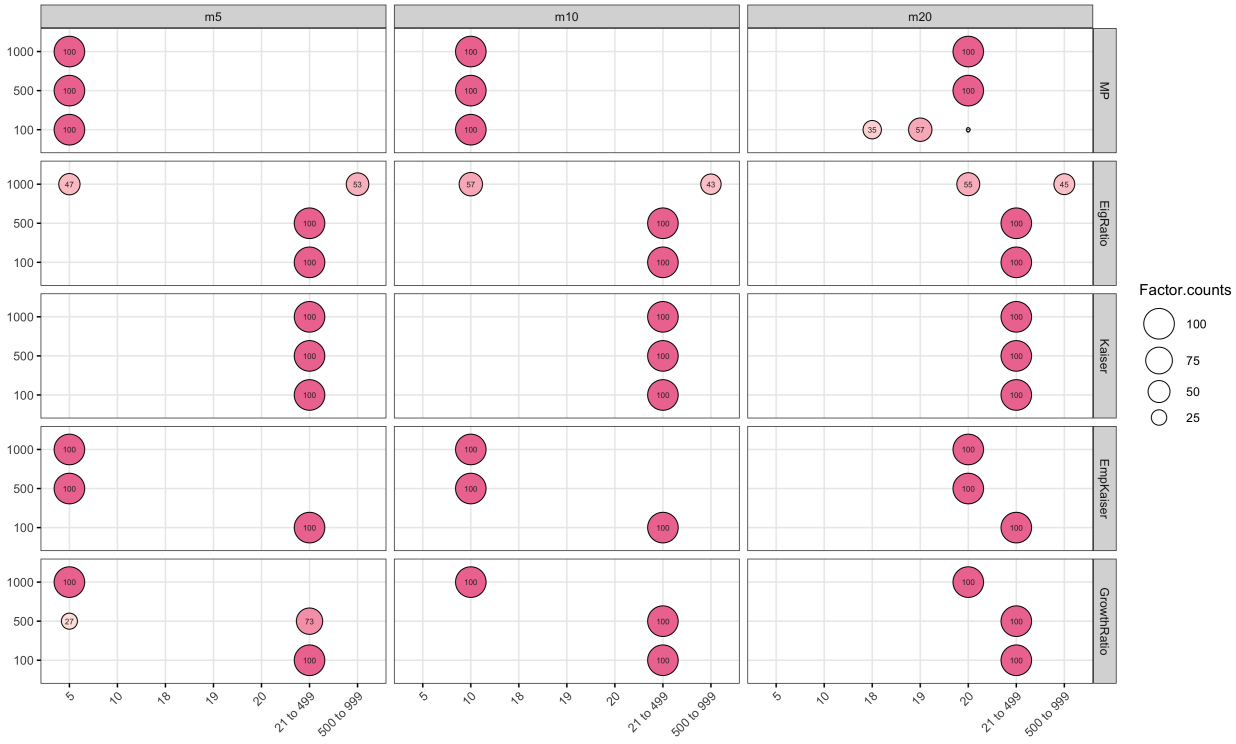


Figure 18: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is of $High$ values from $(.75, .80)$ and λ is of Low values from $(.60, .75)$

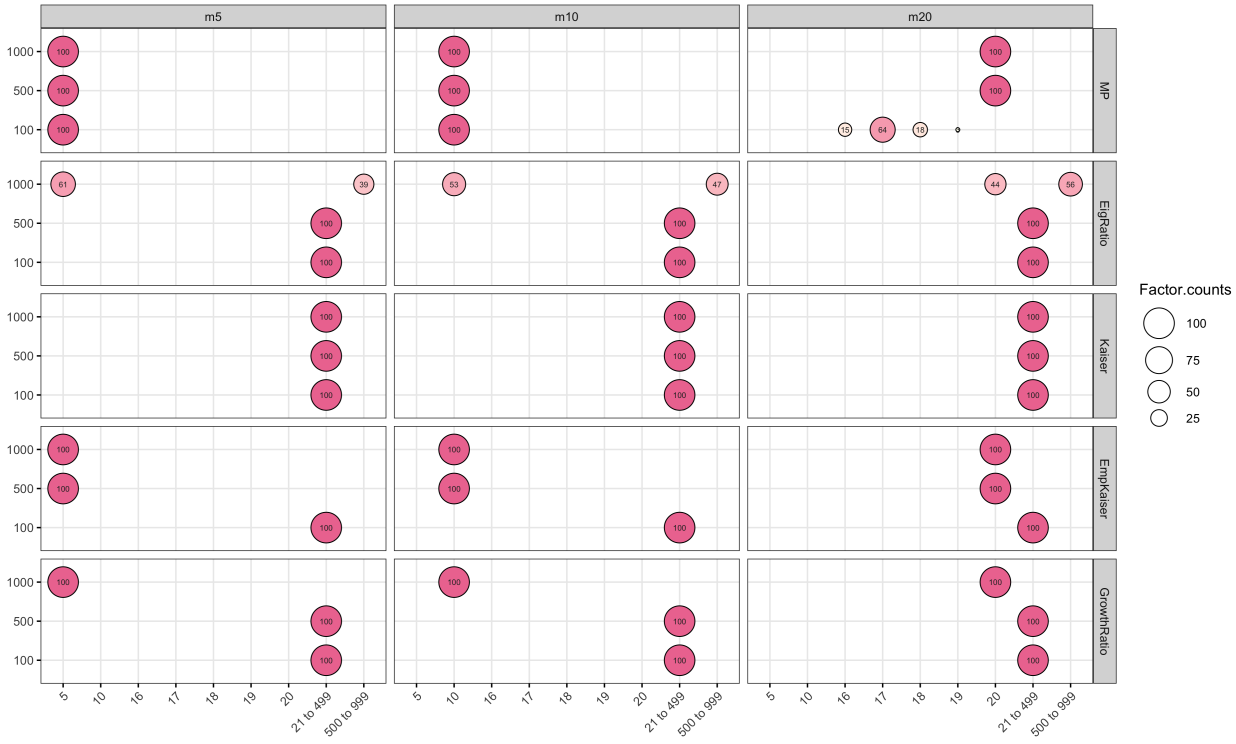
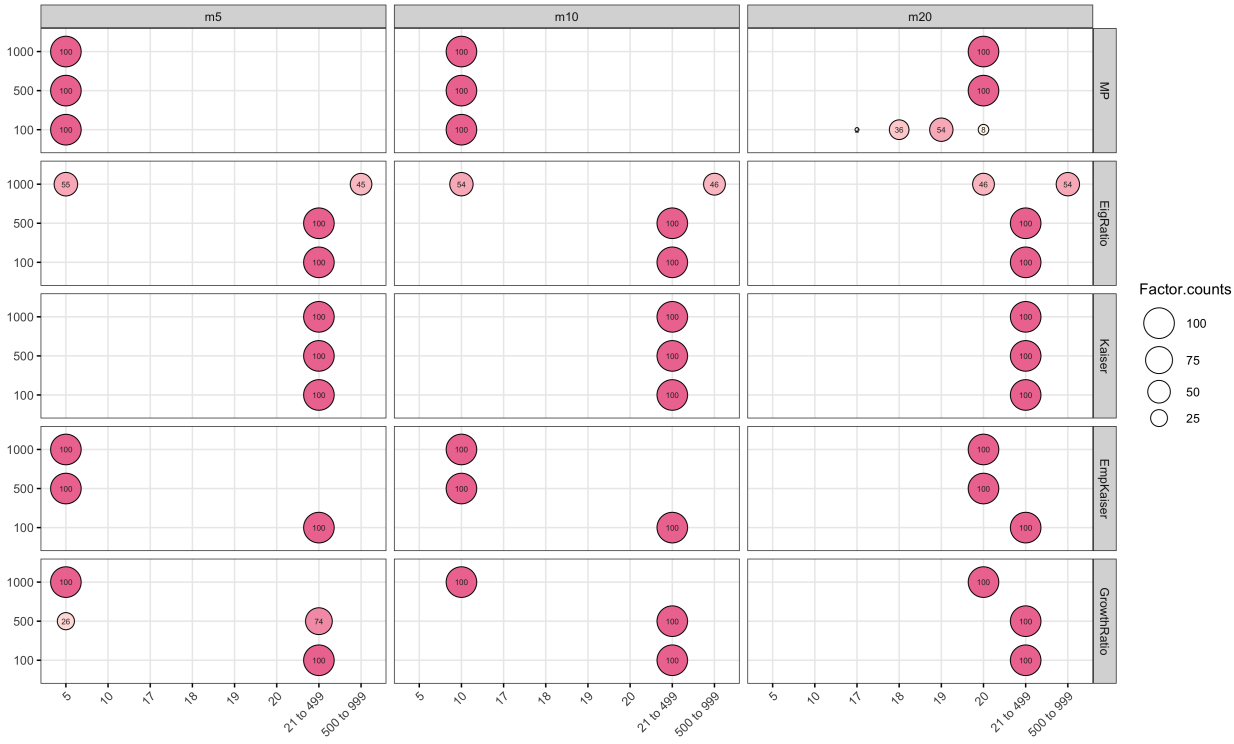


Figure 19: Counted times of estimated m in comparison to the true m for the methods of Marchenko-Pastur law, Eigenvalue ratio, Kaiser's rule, Empirical Kaiser and Growth ratio. $p = 1000$, ϕ is of $High$ values from $(.75, .80)$ and λ is of $High$ values from $(.75, .90)$



Appendix D Factor loadings for small, round blue cell tumors gene expression data

	Latent factors										
	1	2	3	4	5	6	7	8	9	10	11
GENE247		0.52									
GENE524		0.52									
GENE1170		0.50									
GENE1845		0.51									
GENE1	0.33				0.32						
GENE2						0.35					
GENE3											
GENE5											
GENE6											
GENE7											
GENE8							0.41				
GENE9											
GENE10			0.32								
GENE11											

GENE12			0.35						
GENE13									
GENE14									
GENE15		-0.32							
GENE16									
GENE17									
GENE18									
GENE19									
GENE20									
GENE21								0.30	
GENE22									
GENE23									
GENE24									
GENE25			0.34						
GENE26				-0.36					
GENE27									
GENE28		-0.32							
GENE29									
GENE30									
GENE31									
GENE32									
GENE33	0.46								
GENE34									
GENE35									
GENE36									
GENE37									
GENE38									
GENE39									0.32
GENE40									
GENE41									
GENE42									
GENE43		0.39							
GENE44			0.31						
GENE45									
GENE46									
GENE47									
GENE48			0.32						
GENE49									
GENE50	0.33								
GENE51	-0.32								
GENE52									
GENE53									

GENE54						
GENE55			0.31			
GENE56	-0.32					
GENE57						
GENE58		0.41				
GENE59						
GENE60						
GENE61						
GENE62						
GENE63						
GENE64						
GENE65						
GENE66						
GENE67						0.35
GENE68						
GENE69						
GENE70		0.41				
GENE71						
GENE72						
GENE73						
GENE74						
GENE75	0.34					
GENE76	-0.38					
GENE77						
GENE78	0.30					
GENE79						
GENE80				0.35		
GENE81			0.31			
GENE82						
GENE83						
GENE84						
GENE85					-0.39	
GENE86		0.35				
GENE87		-0.35				
GENE88		0.42				
GENE89	0.36					
GENE90						
GENE91						
GENE92						
GENE93						
GENE94						
GENE95						

GENE96							
GENE97							
GENE98							
GENE99					0.30		
GENE100		0.32					
GENE101				0.34			
GENE102							
GENE103							
GENE104							
GENE105							
GENE106							
GENE107	0.31						
GENE108	0.41						
GENE109							
GENE110							
GENE111							
GENE112							
GENE113							
GENE114	0.37						
GENE115		0.41					
GENE116							
GENE117		0.42					
GENE118							
GENE119	0.39						
GENE120							

Table 6: SRBCTs data: factor loadings when $m = 11$. Factor loadings bigger than 0.30 are reported. There are in total 2308 genes, and 2184 genes are omitted. The first four genes in the table are the most significant ones and their loadings are exceeding 0.50

Appendix E Factor loadings for head and neck squamous cell carcinoma data

	Latent factors						
	1	2	3	4	5	6	7
Morphology.3	-0.75			0.45			
Morphology.11	0.69					-0.35	
Morphology.12	0.85						
Morphology.13	0.82				0.36		
glcmFeatures2Dmrg.14	0.65						
glcmFeatures2Dmrg.18	0.72			-0.40			

glcmFeatures2Dvmrg.14	0.82						
glcmFeatures2Dvmrg.16	0.80						
glcmFeatures2Dvmrg.18	0.73					0.50	
glcmFeatures3DWmrg.14	0.84						
glcmFeatures3DWmrg.16	0.82						
glcmFeatures3DWmrg.18	0.81					0.33	
glcmFeatures3DWmrg.24	0.71	0.31				0.39	
GLRLMFeatures2Dvmrg.13	0.61	0.56					0.31
GLRLMFeatures3Dmrg.10	0.83						
GLSZMFeatures2Davg.10	0.76	0.46					
GLSZMFeatures2Dvmrg.7	0.60	0.45			0.41		0.33
GLSZMFeatures3D.10	0.69	0.55					
GLSZMFeatures3D.15	0.54	0.41	-0.44	-0.50			
ngtdmFeatures2avg	-0.52	-0.38			0.45		
gldzmFeatures2Davg.8	0.72	-0.34			0.46		
gldzmFeatures2Davg.10	0.83						
gldzmFeatures2Davg.11	-0.60			0.35			
gldzmFeatures2Davg.14	0.64	0.47	-0.30	-0.42			
gldzmFeatures2Dmrg.8	0.69				0.61		
gldzmFeatures2Dmrg.10	0.85				0.35		
gldzmFeatures2Dmrg.11	-0.64		0.34	0.43			
gldzmFeatures2Dmrg.14	0.78						
gldzmFeatures3D.10	0.84	0.33					
gldzmFeatures3D.14	0.60	0.48					
ngldmFeatures2Davg.10	0.83	0.33					
ngldmFeatures2Davg.15	0.78			-0.41			
ngldmFeatures2Dmrg.14	0.66	0.47		-0.36			
ngldmFeatures3D.4	-0.65	-0.31		0.39			
ngldmFeatures3D.7	0.70	0.42		0.39			0.38
ngldmFeatures3D.10	0.85						
ngldmFeatures3D.15	-0.65	-0.35		0.54			
Statistics.4	0.35	0.83					
Statistics.5		0.77					
Intensity.histogram.8	0.34	0.54					
Intensity.histogram.20	0.31	0.68					
Intensity.histogram.22	0.43	0.66					
intensity.volume.3	0.48	0.74					
glcmFeatures2Davg.6	0.42	0.67		-0.40			
glcmFeatures2Davg.23		-0.78					
glcmFeatures2Dmrg.12		0.90					
glcmFeatures2Dmrg.15		-0.65	0.49	0.46			
glcmFeatures2Dmrg.17		-0.84					

glcmFeatures2Dmrg.21		-0.61				
glcmFeatures2Dmrg.23		-0.76		0.40		
glcmFeatures3Davg.23	0.37	-0.62				0.33
glcmFeatures3Davg.24		-0.69		-0.34		-0.34
glcmFeatures3DWmrg.6		0.79	-0.39	-0.35		
glcmFeatures3DWmrg.11		0.90				
glcmFeatures3DWmrg.17		-0.82		0.33		
glcmFeatures3DWmrg.22		0.63				
GLRLMFeatures3Dmrg.15	0.45	0.73		-0.38		
GLSZMFeatures2Davg		0.57	-0.50	-0.48		
GLSZMFeatures2Davg.11		0.80				
GLSZMFeatures3D		0.72	-0.30	-0.41		
GLSZMFeatures3D.11		0.87				
ngtdmFeatures2avg.4		0.89				
ngtdmFeatures3D.1		0.87				
ngtdmFeatures3D.4		0.83				
gldzmFeatures2Dmrg.7	0.39	0.72				
gldzmFeatures3D.11	-0.47	-0.56		0.31		
ngldmFeatures2Dmrg		0.70	-0.44	-0.39		
ngldmFeatures2Dmrg.7	0.54	0.62				0.40
ngldmFeatures2Dmrg.11		0.81				
ngldmFeatures3D		0.87				
ngldmFeatures3D.5		0.84				
ngldmFeatures3D.11		0.82				
Morphology.9		-0.37	0.64			
glcmFeatures3DWmrg.10			0.75	0.51		
glcmFeatures3DWmrg.23	-0.49		-0.68			
GLRLMFeatures2DWmrg.1			0.89			
GLRLMFeatures2DWmrg.12		0.44	-0.58	-0.42		
GLRLMFeatures2Dvmrg.6			0.91			
GLRLMFeatures3Davg.6			0.83			
GLRLMFeatures3Davg.14			0.89			
GLRLMFeatures3Dmrg.12		0.45	-0.77	-0.32		
GLSZMFeatures2Dvmrg.14			0.79		0.44	
GLSZMFeatures3D.6			0.92			
ngtdmFeatures2avg.2			0.82			
ngtdmFeatures3D.2			0.89			
ngldmFeatures2Dmrg.13		-0.50	0.53	0.34		
ngldmFeatures3D.2			0.64	0.63		
ngldmFeatures3D.9		-0.39	0.63	0.55		
ngldmFeatures3D.13		-0.34	0.77			
intensity.volume.4	0.33		-0.49	-0.64		

glcmFeatures2Davg.10		0.44	0.72				
glcmFeatures2Davg.24	0.33	0.32	-0.32	-0.69			
glcmFeatures2Dmrg			0.54	0.67			
glcmFeatures2Dmrg.24	0.45	0.48		-0.64			
GLRLMFeatures2DWmrg.4	-0.30		0.40	0.73			
GLSZMFeatures2Dvmrg.4	-0.42	-0.39		0.67			
ngtdmFeatures3D	-0.57			0.65			
gldzmFeatures3D.6		-0.33	0.39	0.69			
ngldmFeatures2Davg.4	-0.52	-0.41		0.56			
ngldmFeatures2Davg.16	-0.48			0.62			
ngldmFeatures2Dmrg.15	-0.50	-0.35	0.37	0.55			
GLRLMFeatures3Dmrg.8	0.43				0.82		
GLSZMFeatures3D.1			0.53		0.75		
GLSZMFeatures3D.7					0.87		
ngldmFeatures2Davg.8	0.38		0.42		0.72		
Morphology.15						0.56	
glcmFeatures2Dvmrg.23			-0.54				-0.57
glcmFeatures2Dvmrg.24	0.46	0.43				0.54	-0.37
Intensity.histogram.2							-0.50
Intensity.histogram.14	-0.47						-0.59
Morphology.5	-0.49					0.39	
Morphology.8	0.46					-0.37	
Morphology.14						0.48	
Morphology.17				0.38		-0.46	
Morphology.18			-0.34	-0.46			
Statistics.14		0.43		-0.41			-0.43
Statistics.15		0.43		-0.38			-0.37
Intensity.histogram.3							
Intensity.histogram.15	-0.39						-0.46
intensity.volume.1			0.30	0.45			
glcmFeatures3DWmrg.21							-0.34
GLSZMFeatures3D.4	-0.39	-0.47		0.45			
ngtdmFeatures2avg.1		0.47					
ngldmFeatures2Davg.12	0.46						

Table 7: HNSCC data: factor loadings when $m = 7$. Factor loadings bigger than 0.30 are reported. There are in total 124 features retained after filtering with cutoff value of 0.95