

Article

Comparison of Methods for Addressing Outliers in Exploratory Factor Analysis and Impact on Accuracy of Determining the Number of Factors

W. Holmes Finch

Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA; whfinch@bsu.edu

Abstract: Exploratory factor analysis (EFA) is a very common tool used in the social sciences to identify the underlying latent structure for a set of observed measurements. A primary component of EFA practice is determining the number of factors to retain, given the sample data. A variety of methods are available for this purpose, including parallel analysis, minimum average partial, and the Chi-square difference test. Research has shown that the presence of outliers among the indicator variables can have a deleterious impact on the performance of these methods for determining the number of factors to retain. The purpose of the current simulation study was to compare the performance of several methods for dealing with outliers combined with multiple techniques for determining the number of factors to retain. Results showed that using correlation matrices produced by either the percentage bend or heavy-tailed Student's *t*-distribution, coupled with either parallel analysis or the minimum average partial yield, were most accurate in terms of identifying the number of factors to retain. Implications of these findings for practice are discussed.

Keywords: factor analysis; outliers; robust statistics



Citation: Finch, W.H. Comparison of Methods for Addressing Outliers in Exploratory Factor Analysis and Impact on Accuracy of Determining the Number of Factors. *Stats* **2024**, *7*, 842–862. <https://doi.org/10.3390/stats7030051>

Academic Editor: Wei Zhu

Received: 18 June 2024

Revised: 18 July 2024

Accepted: 25 July 2024

Published: 5 August 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Exploratory factor analysis (EFA) is widely used by researchers in a variety of fields to identify the latent structure underlying a set of observed variables, such as item responses on psychological and educational measurements, subscale scores from test batteries, and measures on observational behavior inventories, as examples. Because EFA does not posit a specific factor structure, a key component of using the methodology involves determining the number of factors to retain. A variety of statistical tools exist for this purpose, which can be used in conjunction with an understanding of the theory underpinning the observed variables to determine the optimal number of factors to retain. Once the number of factors has been decided upon, the researcher can then go on to interpret the nature of the latent variables based on how the observed indicators group together.

Prior research has found that the presence of outliers in the observed data can have a deleterious impact on the accurate determination of the number of factors to retain [1–13]. Specifically, outliers have been found to result in a tendency for researchers to retain more factors than are actually present in the data and a single factor that accounts for a greater share of variance in the sample data than is the case for the population [6]. Given that outliers are common in social science data [14], researchers using EFA may reach incorrect conclusions regarding the latent structure underlying the observed data. Therefore, the availability of alternative approaches for fitting EFA models in the presence of outliers represents an important tool in a researcher's toolbox. The goal of this simulation study was to investigate the performance of several methods for dealing with outliers and to examine how they interact with methods for determining the number of factors to retain under a variety of conditions. The manuscript is organized as follows: First, the common factor model is briefly reviewed, after which there is a discussion of several statistical approaches for determining the number of factors to retain and robust methods for handling outliers.

The simulation study conditions are then described, followed by the results and a discussion of these results with recommendations for practice.

1.1. The Common Factor Model

The common factor model relates the individual observed indicator variables with the latent factors as in Equation (1):

$$y = \Lambda\zeta + \varepsilon \quad (1)$$

where

$y = A (p \times 1)$ vector of observed indicators for the individuals in the sample

$\zeta = A (k \times 1)$ vector of latent factors for the individuals in the sample

$\Lambda = A (p \times k)$ matrix of factor loading matrix linking observed indicators with factors

$\varepsilon = A (p \times 1)$ vector of unique factors for indicator y for the individuals in the sample, i.e., sources of indicator variance that are not associated with ζ .

Note that the model in Equation (1) is for mean-centered variables. The value of the observed indicator is a function of an individual's level on the latent variable(s), the relationship between the latent and observed variables, and the random error associated with the indicator. The ε for indicator y is random and independent of the ε for all other indicators and of ζ . Within the matrix Λ are separate factor loadings relating each indicator to each factor.

The factor model parameters can be used to predict the covariance matrix of the indicator variables:

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \quad (2)$$

where

Σ = Model predicted correlation or covariance matrix of the indicators

Ψ = Correlation matrix for the factors

Θ = Diagonal matrix of unique error variances.

Fit of the factor model is determined by comparing the model implied covariance matrix, Σ , with the observed covariance matrix S . The more similar Σ and S are to one another, the better the fit of the model to the data. On the other hand, if the corresponding elements of the two matrices are far apart, we would conclude that the model does not fit the data well. The proximity of Σ and S can be assessed using a Chi-square statistic, as described in more detail below.

1.2. Maximum Likelihood

One of the most popular and widely used methods of factor extraction is maximum likelihood estimation (ML). The goal of ML in the context of EFA is to identify factor model parameter estimates for Equation (2) that will yield Σ that is as close to S as possible. The mechanism by which the estimated covariance matrix is determined involves minimizing the fit function in Equation (3):

$$F_{ML} = \ln|\Sigma| + \text{tr}(\Sigma^{-1}S) - \ln|S| - p \quad (3)$$

where

Σ = Predicted correlation matrix for the observed variables based on the factor analysis model parameters

$|\Sigma|$ = Determinant of Σ

S = Observed correlation matrix for the observed variables

$|S|$ = Determinant of S

p = Number of observed indicators.

ML uses an iterative methodology to obtain the factor loadings, factor variances, and covariances, which are, in turn, used to calculate Σ using Equation (2). It has been shown to

be an effective method for extracting an initial factor solution, provided that the assumption of multivariate normality is met [15]. Despite its generally positive performance under a variety of conditions, ML is sensitive to the sample size such that with small N , the algorithm may have difficulty converging. In addition, it relies on an assumption of multivariate normality for the observed indicator variables. When the normality assumption is violated, the data analyst should use an alternative estimator, such as ML, with robust standard errors [11,16]. Furthermore, although the focus of this study is on continuous indicator variables, alternative estimators for categorical indicators are also available, such as weighted least squares and diagonally weighted least squares [17,18].

1.3. Methods for Determining the Number of Factors to Retain

There exist a large number of methods for determining the number of factors to retain with EFA. The current study examined the performance of eight such statistics in the simulations described below. The results of the simulation study identified two methods that appeared to perform best across conditions. Therefore, those methods, along with the Chi-square hypothesis test, will be described in more detail below. The full list of techniques included in the simulation study appears in the Methods Section. The Chi-square statistic is discussed here because it is closely tied to the ML estimation algorithm and is the default assessment method in many widely available software programs, such as SPSS, R, SAS, and Stata.

1.4. Parallel Analysis

Another inferential approach for determining the number of factors to retain in EFA is parallel analysis (PA; [19]). PA involves the generation of synthetic data that has the same marginal properties (i.e., means and variances) as the actual observed data but which has no underlying latent structure (i.e., 0 factors). A large number (e.g., 1000) of synthetic random datasets are generated, and for each, a principal components analysis model is fit, and the eigenvalues (i.e., first, second, third, etc.) are retained, thereby creating a distribution for each eigenvalue under the null hypothesis of no factor structure is present. The eigenvalues obtained from EFA applied to the observed data are then compared to these distributions to determine the number of factors to retain. A factor is retained if the observed eigenvalue corresponding to it is larger than the 95th percentile of the distribution of null factor eigenvalues generated from the synthetic random data. It should be noted that a comparison of the observed eigenvalue to the mean of the synthetic eigenvalues has also been used in prior research and application [19]. The interested reader can find a more detailed description of EFA in Finch [20]. Prior research has found that PA is effective at identifying the number of factors to retain [21–24]. For this reason, it was included in the current study.

1.5. Minimum Average Partial

The minimum average partial (MAP) is used to determine the number of factors to retain [25]. MAP is based on the mean of the squared partial correlations among the observed indicators after partialing out the influence of the factors. It is based on the following multiple-step procedure:

1. Correlations among the observed variables are calculated and squared, and the squares are then averaged.
2. A principal components analysis (PCA) is fit to the data, and the squared correlations among the observed indicators are calculated and averaged after partialing out the first factor.
3. The average squared correlation among the observed variables is again calculated, this time partialing out the first two factors obtained from the PCA.
4. These steps are repeated for the first $p - 1$ factors, where p is the number of observed indicators.

5. The researcher retains the number of factors corresponding to the minimum average squared partial correlation obtained from steps 1–4.

The minimum average squared partial correlation identified in step 5 corresponds to the number of factors accounting for the maximum amount of systematic variance in the observed indicators [25]. MAP has been shown to be effective in correctly determining the number of factors to retain [26–29].

1.6. Exploratory Graph Analysis

Recently, authors have proposed the use of an alternative network-based approach to modeling data that has traditionally been addressed using factor analysis [30–32]. This technique, known as exploratory graph analysis (EGA), is based on the Gaussian graphical model (GGM) described in [31]. Golino et al. [32] showed that the GGM and the factor model correspond closely to one another. In particular, the inverse of the covariance matrix for the observed indicator variables (Σ) is known as the precision matrix K such that:

$$K = \Sigma^{-1} \quad (4)$$

The negative elements of K (k_{ij}) can be standardized based on the diagonal elements (k_{jj}, k_{ii}) in order to obtain partial correlation coefficients between any pair of observed variables, i and j :

$$\rho_{ij} = -\frac{k_{ij}}{\sqrt{k_{jj}}\sqrt{k_{ii}}} \quad (5)$$

These partial correlations then serve as the degree of relationship between pairs of variables and are used to identify variable clusters in EGA.

The network-reflecting relationships among the variables are then estimated using the GLASSO technique [33], which is a regularization method based on the lasso estimator [34]. Regularized estimators apply a penalty to parameter estimates (e.g., relationships among variables) such that small sample values are assumed to be 0, and only those that are relatively large remain. In the context of estimating latent variable structure, GLASSO is designed to estimate a sparse network, such that weak connections among the observed variables in the sample are penalized to 0, yielding a network that retains only relatively strong connections among the variables. The degree of regularization is governed by the tuning parameter, γ . As discussed in [32], the optimal value of γ for a given sample is associated with the minimum value of an information statistic, such as the extended Bayesian Information Criterion (eBIC). The number of dimensions to retain in the context of EGA is determined using the Walktrap algorithm, details of which can be found in [3,32].

Prior research investigating the performance of EGA in identifying the number of dimensions to retain has found that it works well in a variety of conditions. For example, [31] found that EGA performed comparably to PA and better than MAP in terms of identifying the correct number of factors across a range of sample sizes, number of indicators, and number of factors. In addition, when the interfactor correlation was large (0.7) and there were four factors, EGA was more accurate than PA. Similar simulation results were reported by [32], who also found that EGA performed comparably to or better than PA across a variety of conditions and was generally more accurate than other techniques included in the study, such as Kaiser's greater than 1 eigenvalue rule, the optimal coordinate, and the acceleration factor techniques. These prior results suggest that EGA is a viable alternative to traditional approaches for identifying the number of factors to retain.

1.7. Impact of Outliers on Determining Number of Factors to Retain

Researchers have examined the impact of outliers on the performance of multiple methods for determining the number of factors to retain in the context of both PCA and factor analysis. Based on the results of a simulation study, [8] found that ML extraction coupled with the Chi-square test tended to overfactor in the presence of outliers. They

also found that the use of the minimum covariance determinant (MCD) to obtain the covariance matrix used in the factor extraction yielded factor solutions that were closer to the latent structure used to generate the data than was the case for the standard covariance matrix [35]. It is important to note that this simulation study included a single factor with five indicators and a sample size of 100. [36] conducted a simulation study that found a tendency to overfactor when outliers were present, and the eigenvalue greater than 1 rule was used to determine the number of factors to retain.

Ref. [6] investigated the impact of outliers on the ability of several approaches, including the Chi-square test, PA, and MAP, to correctly identify the number of factors underlying a set of normally distributed indicator variables. They found that the performance of PA and MAP were largely unaffected when the outliers were symmetric in nature. However, the Chi-square test tended to recommend the retention of too many factors. Furthermore, [6] also reported that when the outliers were asymmetric (i.e., the indicator variables were skewed), MAP tended to overfactor, and PA tended to underfactor. The author concluded that while MAP and PA were generally more resistant to outliers than the Chi-square test, their performance was also deleteriously impacted by outliers in some conditions. [7] extended simulation work by [6,36], confirming earlier results that outliers have a deleterious impact on the performance of the Chi-square test and PA for determining the number of factors to retain. These methods were all used in conjunction with maximum likelihood estimation.

1.8. Methods for Dealing with Outliers

The statistics literature includes a large number of approaches for dealing with outliers [37]. In the context of EFA, it has been recommended that researchers faced with outliers use a robust method for estimating the covariance matrix among the observed indicators and then use the result to extract the factors [8]. This stands in contrast to the default approach for extracting factors, which is based on the standard covariance or correlation matrix. The following section of the manuscript provides a description of robust methods for estimating the covariance matrix among the indicators. For the current study, each of these approaches was applied to the simulated data, and the resulting covariance matrices were then used for factor extraction.

1.9. Percentage Bend Covariance Matrix

The percentage bend (PB) approach is designed to reduce the impact of outliers through a weighting system [37,38]. The reader interested in the technical details of the method is referred to [37] for an in-depth discussion. The PB algorithm starts by calculating the distance between the data value and the median for each variable in the set (e.g., each indicator variable used in the factor analysis):

$$W_i = |X_i - M_X| \quad (6)$$

where

X_i = Value of variable X for individual i

M_X = Median for variable X

The most extreme value that lies at a predetermined quantile of the aforementioned distances (e.g., 75th percentile) is then identified and labeled W_X . For each median distance value that is less than this extreme score, a ratio is calculated with the median distance in the numerator and the extreme distance value in the denominator, as in Equation (7).

$$\frac{X_i - M_X}{W_X} \quad (7)$$

The inverse of the resulting ratio is then used to determine a weight for each observation, with those having values greater than W_X receiving a weight of 0. This process is completed for each variable in the dataset, and the covariance/correlation matrix is

calculated using the resulting weighted data. It should be noted that the standardized version of PB is not an estimate of the population correlation value (ρ) for a variable pair, but rather is an estimate of the linear relationship between the two variables. Therefore, while it provides comparable information to that reflected in ρ , it is not a direct estimate of ρ [37].

1.10. Heavy-Tailed Covariance Matrix

The heavy-tailed covariance matrix (Ht) used in this study is based on the work of [39] and involves the application of the generalized hyperbolic skewed Student's t -distribution. This distribution has parameters μ (mean vector), δ (covariance matrix), γ (vector of sample skewness values), and ν (degrees of freedom). The resulting distribution has heavy tails, which accounts for the skewness present in the sample data. By default, in the R software package used in this study (fitHeavyTail; [40,41]), the degrees of freedom is 4, though the data analyst can adjust this as needed. Smaller values for ν yield heavier-tailed distributions, whereas large values yield distributions approaching the standard normal.

1.11. Winsorized Covariance Matrix

The Winsorized (W; [37]) covariance matrix is obtained by replacing a predetermined proportion of individuals at either extreme of the sample with the same proportion of the next most extreme values. For example, if we want to Winsorize 10% of the sample, the extreme 10% members at each end of the sample are replaced by the 10% adjacent to them. In the context of multivariate data (such as that used in EFA), we cannot order the data based on any single value, but instead need to order the individuals based on the full set of variables. Thus, in order to characterize each individual based on the set of variables used in the EFA, the Mahalanobis distance based on these variables is calculated for each member of the sample. The Mahalanobis distance value reflects the location of each individual in multivariate space, with values closer to 0 being near the multivariate center of the observed variables. Multivariate Winsorization is based on these distances, such that individuals with the 10% smallest Mahalanobis values have their full set of variables replaced by those in the next highest 10% based on the Mahalanobis distance. Similarly, the set of variables used in the EFA for individuals with the largest 10% of Mahalanobis distance values is replaced by scores for the 10% of individuals with the next highest Mahalanobis distance values. Thus, individuals are ordered based on the Mahalanobis distance calculated using the set of indicator variables, and then the desired proportion at each end, based on the Mahalanobis distance, is Winsorized. The covariance matrix from the resulting sample is then used for factor extraction.

1.12. Minimum Volume Ellipsoid

The minimum volume ellipsoid (MVE; [42]) involves finding a subsample of the original dataset that creates an ellipsoid with the smallest volume in the multivariate space represented by the set of observed variables and consists of one-half or more of the original sample. As an example, if the total sample size is 100, then the MVE subsample will include between 50 and 100 individuals. The observations contained within this minimum volume ellipsoid set are then used to estimate the covariance matrix that is subsequently used in the EFA. In order to identify the subset of individuals, a large number ($B = 1000$) of random samples with replacement of size $h = \frac{n+p+1}{2}$ are taken from the original sample, where n is the sample size, and p is the number of observed variables. For each of these random samples, the volume of the ellipsoid that contains the entire set of observations is calculated, and the sample that yields the smallest value is selected as optimal. In the context of EFA, the covariance matrix of the optimal MVE sample is then used to extract the factors. Figure 1 displays a plot of MVE for two variables. The data points contained within the ellipse are those that will be retained and used to estimate the covariance. The outliers appear outside of the ellipse with their case number. They will not be included in the estimation of the covariance.

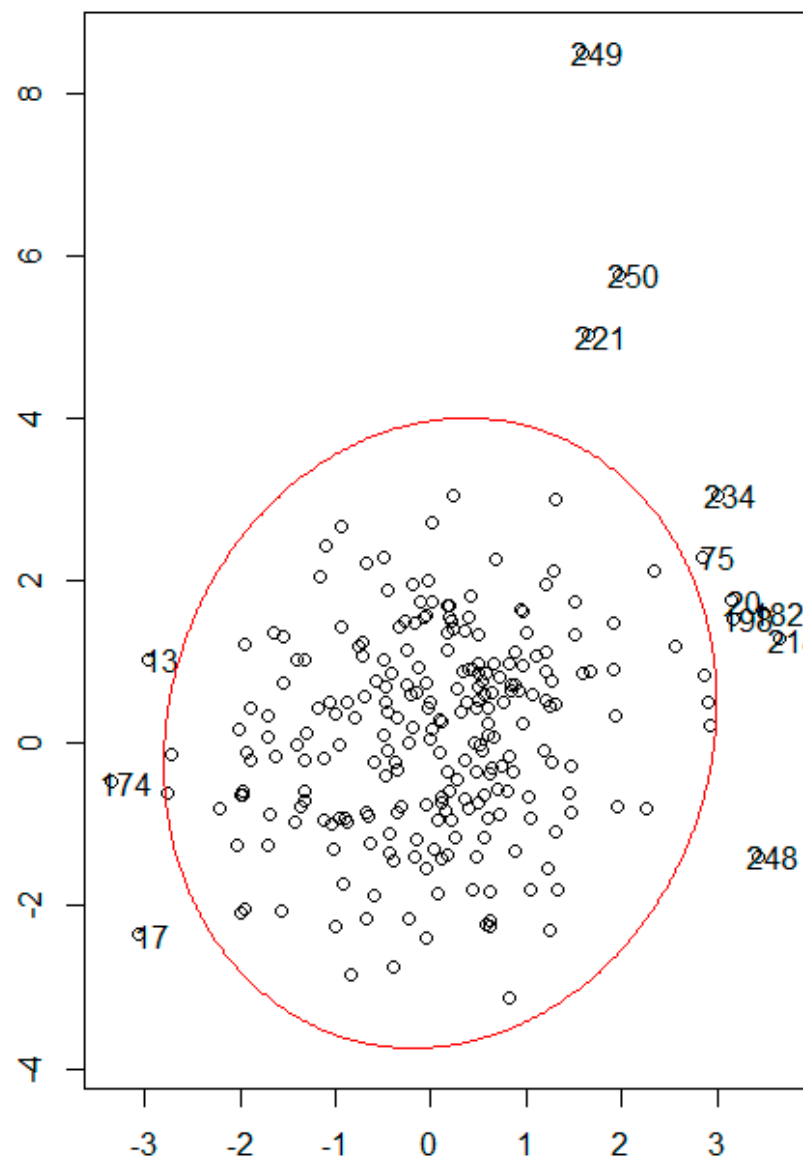


Figure 1. Example minimum volume ellipsoid.

1.13. Minimum Covariance Determinant

The minimum covariance determinant (MCD; [35]) is similar to MVE in that it also involves a systematic search through the data for a subset of observations containing no outliers. However, rather than identifying a set of observations that minimizes the ellipsoid, MCD seeks to identify the set of observations that minimize the determinant of the covariance matrix. The MCD algorithm involves the following steps:

1. Select a subset of size $h = \frac{n+p+1}{2}$ from the original sample.
2. Calculate the Mahalanobis distance from the center for each case in this subset.
3. Retain the $h_2 = \frac{h}{2} + 1$ observations with the smallest distance values.
4. Iteratively include one randomly selected data point from the full sample that was not included in the original subset h_2 .
5. Calculate the determinant of the covariance matrix for this new subset.
6. Repeat steps 4 and 5 and retain the subset that has the smallest determinant of the covariance matrix.

The resulting covariance matrix can then be used to extract the factors with EFA. An example plot for the MCD appears in Figure 2. As with the MVE plot in Figure 1, outliers

appear outside the ellipse and are denoted by their case number. Note that for this example, the MVE and MCD techniques yielded very similar results.

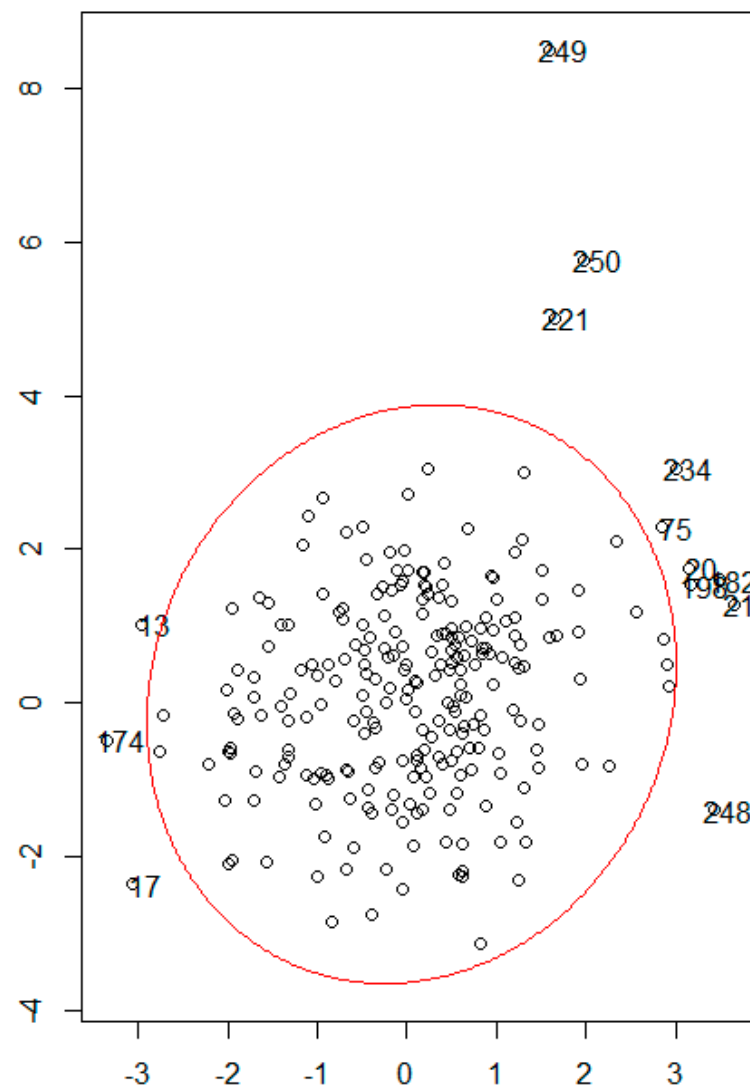


Figure 2. Example minimum covariance determinant.

1.14. MM-Estimator

The *MM*-estimator of the covariance matrix is designed to down-weight outlying observations when estimating the covariance matrix and yields the highest possible breakdown point. Note that the breakdown point refers to the minimum proportion of the sample for which a function goes to infinity that can include outliers without impacting sample estimates [37]. The weight associated with an individual data point is proportional to the inverse of the residual for the individual based on the estimated covariance matrix. These weights are selected such that the sum of squared residuals is minimized across the sample. *MM* has been shown to be effective for high dimensional data in which the ratio of sample size to number of variables is small [37]). However, the *MM*-estimator is known to be susceptible to contamination bias, such that a few outlying data points can have an outside impact on the quality of the parameter estimates, even in the case where the breakdown point approaches 0.5.

1.15. Rocke

Ref. [43] proposed a robust bi-weight estimate of the covariance matrix based on the S-estimator. The original S-estimator [35] takes the form:

$$\frac{1}{n} \sum_{i=1}^n \xi(d_i) \quad (8)$$

where

d_i = Mahalanobis distance value for observation i

$$\xi = \begin{cases} 1 - \left\{ 1 - \left(\frac{d_i}{k} \right)^2 \right\}^3 & \text{if } |d_i| \leq k \\ \xi = 1 & \text{if } |d_i| > k \end{cases}$$

$$k = \sqrt{p \left[\sqrt{\left(\frac{1}{9} \right) \left(\frac{2}{p} \right) 1.548} + \left\{ 1 - \left(\frac{1}{9} \right) \left(\frac{2}{p} \right) \right\} \right]^3}$$

p = Number of variables

The S-estimator is designed to down-weight observations that have large Mahalanobis distance values, where large is defined by the function in Equation (6). It was found that the S-estimator does not always correctly identify outliers [43], which led to the development of Rocke's extension of it. The difference between the Rocke and S-estimator is in the way that outliers are identified and weighted. The S-estimator weight function ξ takes the following form in Rocke's technique:

$$\begin{aligned} \xi &= 0 \text{ if } 0 \leq |d_i| \leq 1 - \gamma \\ \xi &= \left\{ \left(\frac{d_i - 1}{4\gamma} \right) \left[3 - \left(\frac{d_i - 1}{\gamma} \right)^2 \right] + 0.5 \right\} \text{ if } 1 - \gamma < |d_i| < 1 + \gamma \\ \xi &= 1 \text{ if } |d_i| > 1 + \gamma \end{aligned} \quad (9)$$

where

$$\gamma = \min \left(\frac{\chi_p^2(1 - \alpha)}{p} - 1, 1 \right)$$

$\chi_p^2(1 - \alpha)$ = Chi-square value for p degrees of freedom and $1 - \alpha$ probability level.

Rocke's estimator was shown to yield a breakdown point near 0.5 while at the same time correctly identifying outliers in a dataset. For more details on the specific calculations involved in this estimate, the interested reader is referred to [35,44].

1.16. Study Goals

The goal of this simulation study was to extend earlier work by comparing the performance of several methods for dealing with outliers coupled with several approaches for identifying the number of factors to retain. Earlier research described above found that outliers can have a deleterious impact on methods for detecting the number of factors to retain in an EFA (e.g., [7]). However, methods for dealing with outliers were not widely explored in these studies. Therefore, the current study extends this work by assessing the extent to which several methods that account for outliers in the estimation of covariance/correlation matrices can ameliorate the aforementioned problems associated with determining the number of factors in EFA when outliers are present. In addition, the performance of a promising new approach to characterizing the structure underlying a set of observed variables, EGA, has not been investigated in the presence of outliers. Therefore, the current study examined the performance of EGA in the context of outliers among the observed variables. The primary outcome of interest was the proportion of correct instances of the number of retained factors, with a secondary goal of the mean number of factors retained by each combination of outlier methods and techniques for identifying the number of factors

to retain. Based on the prior research discussed above, it is hypothesized that the robust methods for estimating the correlation matrix will yield more accurate results regarding the number of factors to retain when compared with the standard approach. In addition, it is hypothesized that in conjunction with the robust correlation matrix estimation approach, PA will be the most accurate technique for identifying the number of factors to retain.

2. Materials and Methods

The study goals described above were addressed using a simulation study. For each combination of simulated conditions, which are described below, 1000 replications were generated. Data generation and analyses were all conducted using R version 4.1. A four-factor model was simulated with the indicators being generated from the standard normal distribution having a mean of 0 and a standard deviation of 1. The factor loadings were set to 0.8, the factor variances to 1, and the factor correlations to 0.4. Data were generated using the lavaan package [45] from R. The manipulated study conditions are described below.

2.1. Factor Loading Values

Three factor loading conditions were included in this study: 0.4, 0.6, and 0.8. These values were drawn from prior research [6,24] and represent small, medium, and large relationships between the observed indicators and latent variables.

2.2. Number of Indicators per Factor

The number of indicators per factor for each of the four factors was 3, 6, or 12, yielding a total number of indicators of 12, 24, or 48. For a given number of indicators, each factor had the same number of indicators, e.g., 3 indicators for each of 4 factors, yielding a total of 12 indicators.

2.3. Interfactor Correlation

The interfactor correlations were simulated to be 0.2, 0.4, or 0.8, reflecting small, medium, and large relationships among the latent variables [31]. These values have been used in previous research investigating the performance of exploratory factor analysis [6,24].

2.4. Proportion of Sample That Was Outliers

The proportion of individuals in the sample contaminated with outliers was 0, 0.01, 0.08, and 0.15. These conditions represent a range of situations, from no outliers in the sample to a relatively large share, and were borrowed from the work of [6].

2.5. Number of Variables Exhibiting Outliers

A range of values for the number of variables containing outliers was used in the study, including 1, 6, 12, and 24. These values were included to allow for examination of the case when outliers were a minor issue (1), a moderate issue (6 and 12), and when all variables included outliers (24).

2.6. Mean and Standard Deviation Shift

Multiple mean and standard deviation shift for outlier conditions were included in the study and completely crossed with one another. For the mean shift condition, variables for the sample members with outliers were 0 (the means were the same for the main sample and the outlier sample), 1.5, and 3. The standard deviation shift was 1 (the standard deviations for the general sample and the outlier sample were the same), 1.5, and 3. Values for both the mean shift and standard deviation shift were borrowed from earlier research [6].

2.7. Sample Size

The sample sizes used in the study were 250, 500, and 1000 and corresponded to those used in [6]. They represent samples ranging from relatively small (250) to relatively large (1000) in the context of EFA practice.

2.8. Methods for Determining the Number of Factors

In order to determine the number of factors to be retained, multiple methods were used for each simulation replication, including PA, MAP, and EGA. PA and MAP were obtained using the R psych package [46] and EGA from the EGAnet R package [47]. With respect to PA, principal axis factor analysis was used to extract the factors, and the 95th percentile of simulated eigenvalues was used as the cut-off against which the observed eigenvalues were compared. The PA comparison data were generated using resampling of the sampled observations with 1000 datasets. Principal axis factoring was applied to each of these randomly sorted datasets, and the resulting eigenvalues were retained to create the comparison distribution. For each simulated dataset, each method was applied, and the resulting number of factors to retain was recorded. These were then used to calculate the study outcomes, which are described below.

2.9. Methods for Dealing with Outliers

The methods used to address the presence of outliers were Percentage Bend (PB), Winsorized covariance matrix (W), minimum volume ellipsoid (MVE), minimum covariance determinant (MCD), Rocke estimator, MM-estimator, and the heavy-tailed skewed *t*-distribution (Ht). PB and W were carried out using the WRS2 [48] R package, ROCKE and MM with the RobStatTM package [49], and MVE and MCD with the cov.rob function in the MASS package. For each simulation replication, each of these methods was used to estimate the covariance matrix for the observed indicator variables. The resulting covariance matrices were then used for EFA factor extraction, and the techniques for determining the number of factors were used.

2.10. Study Outcomes

The primary outcome of this study was the proportion of replications for each combination of conditions that correctly identified the number of factors to retain (4). The secondary study outcome was the mean number of factors that were recommended for retention across replications. In order to identify the manipulated results and their interactions that impacted the primary study outcome, a mixed-effects analysis of variance (ANOVA) was used in conjunction with the partial ω^2 effect size. The within-subject effect was the combination of dimensionality and outlier detection method (6 levels), and the between-subject effects included the other manipulated factors. For each combination of conditions, the proportion of replications across the 1000 was calculated and then served as the dependent variable for the ANOVA. The assumptions of normality and heteroscedasticity were assessed and found to be met. The ANOVA was conducted using SPSS version 27.

3. Results

The results of the simulation study revealed that the Ht and PB approaches yielded the most accurate results with respect to the number of factors to be retained across all conditions. Therefore, in order to ensure that the results are understandable, clear, and as parsimonious as possible, the following discussion will focus on the simulation results for the following combinations of outlier and factor determination methods: PB/PA, PB/MAP, PB/EGA, Ht/PA, Ht/MAP, and Ht/EGA. In addition, the results for EFA extraction using the standard covariance matrix (S) are also included in the results because they are the default available in most statistical software packages. The results for all of the extraction and covariance estimation methods are presented in Tables 1–8.

Table 1. Proportion of correct number of factors by method, outlier approach, number of variables containing outliers, and proportion of sample with outliers.

V *	C	S EGA	PB EGA	Ht EGA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
1	0.01	0.91	0.94	0.93	0.93	0.93	0.93	0.90	0.90	0.90
	0.08	0.91	0.93	0.92	0.93	0.93	0.93	0.90	0.90	0.90
	0.15	0.90	0.93	0.94	0.93	0.93	0.93	0.90	0.90	0.90
6	0.01	0.92	0.94	0.94	0.92	0.93	0.94	0.89	0.90	0.90
	0.08	0.92	0.94	0.94	0.94	0.94	0.95	0.88	0.90	0.90
	0.15	0.90	0.94	0.92	0.93	0.92	0.95	0.89	0.88	0.88
12	0.01	0.91	0.93	0.93	0.92	0.94	0.95	0.86	0.89	0.88
	0.08	0.86	0.92	0.90	0.86	0.93	0.95	0.81	0.88	0.88
	0.15	0.84	0.90	0.90	0.83	0.92	0.93	0.70	0.87	0.87
24	0.01	0.82	0.93	0.94	0.78	0.93	0.93	0.62	0.85	0.86
	0.08	0.67	0.92	0.92	0.61	0.91	0.92	0.46	0.84	0.84
	0.15	0.59	0.85	0.90	0.59	0.86	0.91	0.40	0.83	0.84

* V = Number of contaminated variables; C = Proportion of contaminated sample; S = Standard estimator using full covariance matrix; PB = Percentage bend correlation; Ht = Heavy-tailed distribution; EGA = Exploratory graph analysis; PA = Parallel analysis; MAP = Minimum average partial.

Table 2. Proportion of correct number of factors by method, outlier approach, number of variables containing outliers, and shift in standard deviation for cases with outliers.

V *	S	S EGA	PB EGA	Ht EGA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
1	1.0	0.95	0.96	0.94	0.94	0.94	0.95	0.93	0.92	0.93
	1.5	0.92	0.93	0.92	0.92	0.91	0.91	0.91	0.92	0.91
	3.0	0.92	0.92	0.93	0.91	0.91	0.92	0.90	0.91	0.91
6	1.0	0.95	0.95	0.95	0.95	0.94	0.95	0.92	0.93	0.92
	1.5	0.90	0.94	0.93	0.90	0.91	0.92	0.88	0.91	0.91
	3.0	0.88	0.94	0.93	0.87	0.90	0.92	0.84	0.88	0.89
12	1.0	0.94	0.95	0.95	0.94	0.94	0.95	0.92	0.92	0.93
	1.5	0.88	0.93	0.92	0.86	0.90	0.92	0.82	0.87	0.87
	3.0	0.75	0.88	0.90	0.77	0.90	0.91	0.72	0.85	0.85
24	1.0	0.94	0.95	0.94	0.94	0.94	0.94	0.92	0.90	0.91
	1.5	0.87	0.91	0.90	0.88	0.89	0.90	0.80	0.84	0.85
	3.0	0.48	0.86	0.89	0.26	0.85	0.89	0.21	0.79	0.81

* V = Number of contaminated variables; C = Proportion of contaminated sample; S = Standard estimator using full covariance matrix; PB = Percentage bend correlation; Ht = Heavy-tailed distribution; EGA = Exploratory graph analysis; PA = Parallel analysis; MAP = Minimum average partial.

Table 3. Proportion of correct number of factors by method, outlier approach, number of variables containing outliers, and shift in mean for cases with outliers. Accuracy rates of 0.90 or higher are in bold.

V *	M	S EGA	PB EGA	Ht EGA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
1	0	0.95	0.96	0.95	0.95	0.94	0.94	0.93	0.93	0.92
	1.5	0.92	0.93	0.93	0.92	0.91	0.92	0.86	0.90	0.91
	3.0	0.91	0.92	0.93	0.92	0.91	0.91	0.81	0.88	0.90
6	0	0.94	0.95	0.95	0.95	0.95	0.94	0.93	0.93	0.93
	1.5	0.91	0.91	0.92	0.91	0.91	0.91	0.86	0.90	0.90
	3.0	0.89	0.90	0.92	0.90	0.90	0.91	0.82	0.87	0.87
12	0	0.94	0.94	0.95	0.95	0.94	0.95	0.93	0.92	0.92
	1.5	0.88	0.92	0.93	0.89	0.91	0.91	0.79	0.86	0.87
	3.0	0.74	0.90	0.91	0.75	0.91	0.91	0.70	0.84	0.85
24	0	0.95	0.94	0.93	0.94	0.95	0.94	0.92	0.92	0.93
	1.5	0.84	0.90	0.91	0.85	0.90	0.90	0.77	0.84	0.84
	3.0	0.70	0.88	0.89	0.70	0.87	0.87	0.62	0.82	0.81

* V = Number of contaminated variables; C = Proportion of contaminated sample; S = Standard estimator using full covariance matrix; PB = Percentage bend correlation; Ht = Heavy-tailed distribution; EGA = Exploratory graph analysis; PA = Parallel analysis; MAP = Minimum average partial.

Table 4. Proportion of correct number of factors by method, sample size, factor loadings, and factor correlations.

N *	S EGA	PB EGA	Ht EGA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
250	0.88	0.91	0.92	0.87	0.90	0.92	0.86	0.89	0.89
500	0.90	0.94	0.94	0.90	0.93	0.94	0.89	0.91	0.92
1000	0.93	0.95	0.95	0.92	0.94	0.95	0.91	0.92	0.93
L	S EGA	PB GA	Ht GA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
0.4	0.87	0.90	0.92	0.85	0.90	0.91	0.82	0.87	0.88
0.6	0.91	0.94	0.95	0.90	0.93	0.94	0.87	0.90	0.90
0.8	0.93	0.97	0.97	0.93	0.96	0.96	0.91	0.92	0.94
C	S EGA	PB GA	Ht GA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
0.2	0.92	0.94	0.95	0.92	0.93	0.95	0.90	0.92	0.92
0.4	0.90	0.93	0.93	0.90	0.93	0.94	0.88	0.90	0.90
0.8	0.87	0.90	0.91	0.87	0.89	0.90	0.84	0.86	0.88

* N = Sample size; L = Factor loading; C = Factor correlation; S = Standard estimator using full covariance matrix; PB = Percentage bend correlation; Ht = Heavy-tailed distribution; EGA = Exploratory graph analysis; PA = Parallel analysis; MAP = Minimum average partial.

Table 5. Mean number of factors retained by method and standard deviation shift.

S *	S EGA	PB EGA	Ht EGA	S PA	PB PA	Ht PA	S MAP	PB MAP	Ht MAP
1.0	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
1.5	4.02	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
3.0	4.13	4.08	4.02	4.15	4.09	4.03	3.68	3.87	3.92

* S = Standard deviation shift; S = Standard estimator using full covariance matrix; PB = Percentage bend correlation; Ht = Heavy-tailed distribution; EGA = Exploratory graph analysis; PA = Parallel analysis; MAP = Minimum average partial.

Table 6. Mean, standard deviation, minimum, and maximum for off-diagonal elements of correlation matrices: dataset 1 (N = 260)/dataset 2 (N = 107).

Correlation Matrix	Mean	Standard Deviation	Minimum	Maximum
Standard overall	0.20/0.19	0.17/0.16	0.10/−0.01	0.49/0.48
Standard block 1	0.26/0.24	0.15/0.13	0.19/0.09	0.35/0.48
Standard block 2	0.25/0.21	0.14/0.12	0.12/0.002	0.46/0.39
Standard block 3	0.24/0.22	0.14/0.14	0.11/−0.01	0.49/0.49
Standard block 4	0.25/0.18	0.15/0.15	0.10/−0.04	0.41/0.45
Percentage bend overall	0.17/0.16	0.11/0.14	−0.11/−0.14	0.51/0.57
Percentage bend block 1	0.21/0.18	0.11/0.10	−0.01/−0.01	0.51/0.44
Percentage bend block 2	0.25/0.20	0.11/0.10	0.13/−0.03	0.47/0.42
Percentage bend block 3	0.22/0.19	0.12/0.11	0.12/−0.01	0.48/0.50
Percentage bend block 4	0.24/0.20	0.09/0.12	0.13/−0.04	0.40/0.57
Ht overall	0.18/0.16	0.11/0.13	−0.09/−0.11	0.52/0.53
Ht block 1	0.22/0.20	0.11/0.09	0.01/0.12	0.51/0.50
Ht block 2	0.24/0.20	0.11/0.11	0.12/0.09	0.47/0.44
Ht block 3	0.22/0.20	0.12/0.11	0.14/0.06	0.50/0.48
Ht block 4	0.25/0.17	0.10/0.10	0.16/0.03	0.46/0.53

Table 7. Number of factors retained by method: dataset 1 (N = 260)/dataset 2 (N = 107).

Extraction/Outlier Method	Number of Factors Retained
Parallel analysis/Standard	5/5
Parallel analysis/PB	4/2
Parallel analysis/Ht	4/4
MAP/Standard	2/2
MAP/PB	3/2
MAP/Ht	3/3
EGA/Standard	5/5
EGA/PB	4/4
EGA/Ht	4/4

Table 8. Factor analysis results for subscales by outlier method. (Data set 1/Data set 2).

Subscale	ML Standard	ML PB	ML Ht	EGA Standard	EGA PB	EGA Ht
1	1/1	1/1	1/1	4/4	4/4	3/3
2	1/1	1/1	1/1	4/4	4/4	3/3
3	1/1	1/1	1/1	4/4	4/4	3/3
4	1/1	1/1	1/1	4/4	4/4	3/3
5	1/1	1/1	1/1	4/1	4/4	3/3
6	1/1	1/1	1/1	4/4	4/4	3/3
7	2/2	2/2	2/2	1/1	5/5	4/4
8	2/2	2/2	2/2	1/1	5/5	4/4
9	2/2	2/2	2/2	1/1	5/5	4/4
10	2/2	2/2	2/2	1/1	5/5	4/4
11	2/2	2/2	2/2	1/1	5/5	4/4
12	2/2	2/2	2/2	1/1	5/5	4/4
13	5/5	4/4	4/4	2/2	2/2	1/1
14	4/4	4/4	4/4	2/2	2/2	1/1
15	4/4	4/4	4/4	2/2	2/2	1/1
16	4/4	4/4	4/4	2/2	2/2	1/1
17	4/4	4/4	4/4	2/2	2/2	1/1
18	4/4	4/4	4/4	2/2	2/2	1/1
19	3/3	3/3	3/3	3/3	1/1	2/2
20	3/3	3/3	3/3	3/3	1/1	2/2
21	3/3	3/3	3/3	3/3	1/1	2/2
22	3/3	3/3	3/3	3/3	1/1	2/2
23	3/3	3/3	3/3	1/1	1/1	2/2
24	3/3	3/3	3/3	3/3	1/1	2/2

3.1. Accuracy Rate for Correctly Identifying Number of Factors

The ANOVA identified the following terms to be statistically significant with respect to the proportion of cases for which the number of factors was correctly identified: interaction of the number of contaminated variables (V) by the proportion of the sample that is contaminated with outliers (C) by method ($F_{42,450} = 6.13$, $p < 0.001$, $\omega^2 = 0.32$), V by standard deviation shift (S) by method ($F_{42,450} = 6.88$, $p < 0.001$, $\omega^2 = 0.44$), V by mean shift (M) by method ($F_{42,450} = 3.49$, $p < 0.001$, $\omega^2 = 0.22$), and sample size (N) by method ($F_{14,142} = 27.24$, $p < 0.001$, $\omega^2 = 0.76$).

Table 1 includes the proportion of the number of factors correctly retained by the number of contaminated variables (V), the proportion of sample contaminated (C), and the method of identification/method of handling outliers. Note that for each combination of V and C, accuracy rates of 0.90 or above are in bold.

PA and EGA generally yielded the highest accuracy rates across conditions when coupled with covariances matrices using either PB or Ht. The use of the standard covariance matrix (S) was consistently associated with the lowest accuracy rates, except when the number of contaminated variables was one (regardless of proportion contaminated) or six

with a contamination of 0.01. In addition, MAP was somewhat less accurate than either EGA or PA for most simulated conditions.

The accuracy rate by method, number of contaminated variables, and standard deviation shift (S) appear in Table 2. As with Table 1, accuracy rates of 0.90 or higher are in bold.

As was the case in Table 1, the Ht/PA, Ht/EGA, PB/PA, and PB/EGA combinations exhibited the highest accuracy rates across conditions. The standard covariance extraction approach yielded the lowest accuracy rates, except for with one contaminated variable or six contaminated variables accompanied by a standard deviation shift of 1 or 1.5. Likewise, MAP yielded lower accuracy rates than either EGA or PA for each of the outlier methods.

Table 3 includes the accuracy rates by number of contaminated variables, mean shift (M), and combination of outlier handling and factor determination methods.

These results are very similar to those in Table 2, with Ht/PA, Ht/EGA, PB/PA, and PB/EGA yielding the most accurate results and S being the least accurate, except with the lowest levels of contamination. In addition, PB and Ht were more impervious to the number of contaminated variables and the degree of mean shift when compared to S, which exhibited worse performance with respect to identifying the correct number of factors when the number of contaminated factors and/or the degree of mean shift increased in value. With respect to methods for determining the number of factors, MAP was less accurate than either PA or EGA.

The accuracy rates by sample size (N), factor loading value (L), interfactor correlations (C), and method (Table 4) show that Ht/PA, Ht/EGA, PB/PA, and PB/EGA had the highest accuracy rates across sample sizes. Accuracy rates of 0.90 or higher are in bold.

For all of the methods, accuracy improved concomitantly with increases in sample size. Similarly, the methods included in the study yielded higher accuracy rates when the factor loadings were larger. Conversely, across factor extraction and outlier methods, accuracy rates were lower when the interfactor correlations were larger. Finally, the lowest accuracy rates were associated with S across the methods for determining the number of factors, sample sizes, factor loading values, and interfactor correlations.

3.2. Mean Number of Factors Retained

ANOVA was used to identify manipulated factors and their interactions that were associated with the mean number of factors to be retained. The results of this analysis identified the interaction of the number of contaminated variables (V) by the proportion of sample that is contaminated (C) by method ($F_{42,450} = 3.94$, $p < 0.001$, $\omega^2 = 0.19$), and the standard deviation shift (S) by method ($F_{16,306} = 9.29$, $p < 0.001$, $\omega^2 = 0.43$). The mean number of factors identified as optimal, by number of contaminated variables, proportion of contaminated variables, and combination of methods for dealing with outliers and determining the number of factors to retain, appear in Figure 3.

The impact of the standard deviation shift on the performance of the extraction methods in terms of the number of factors to be retained appears in Table 5.

When the standard deviation shift was three, the standard covariance matrix coupled with either EGA or PA yielded an inflated number of factors to retain. Inflation in the number of factors retained was also present for EGA with PB and Ht, with the least such effect for Ht/EGA and Ht/PA. In contrast, when the standard deviation shift was three, MAP was associated with underfactored solutions, particularly for S/MAP.

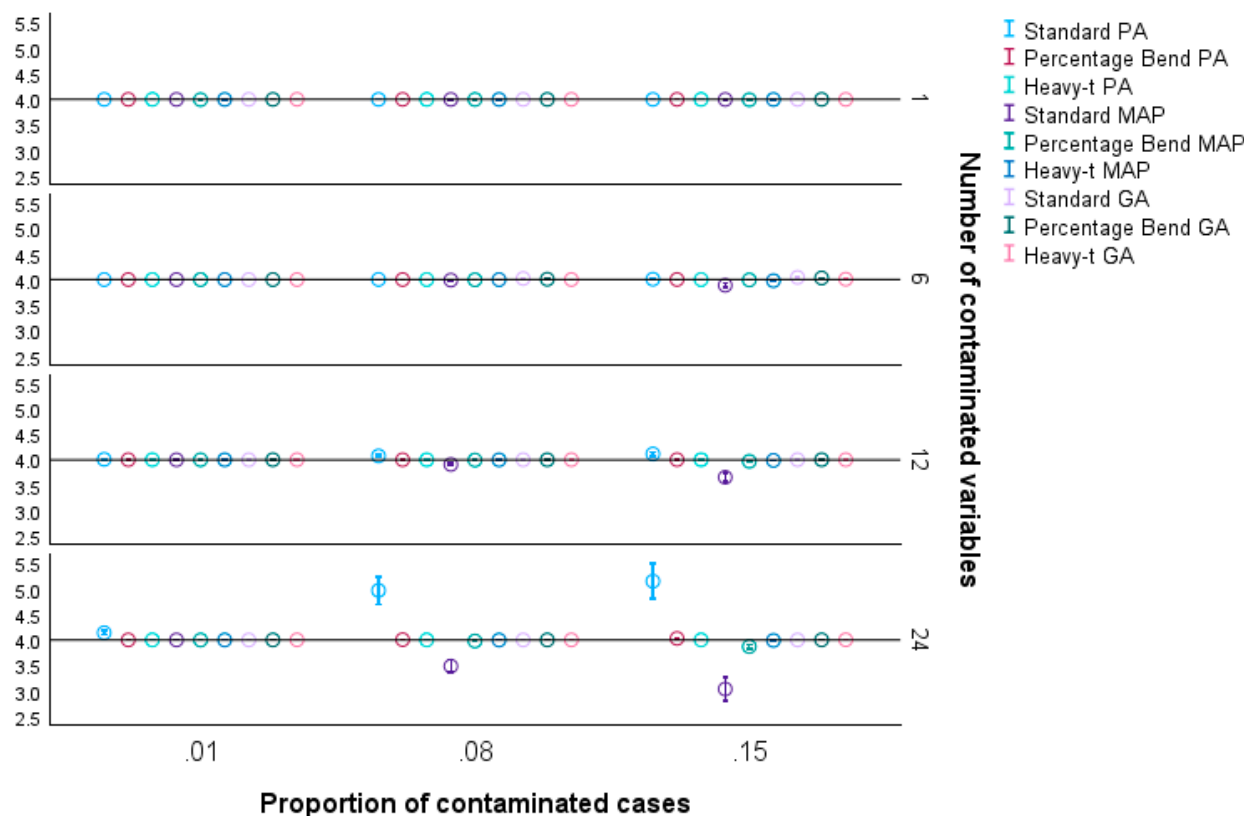


Figure 3. Mean number of factors retained by method, number of contaminated variables, and proportion of sample that is contaminated with reference line at 4 factors. *Reference line at 4 factors. When the number of contaminated variables was 1 or 6, all of the methods indicated that the correct number (4) variables should be retained. As the number of contaminated variables and proportion of the sample that was contaminated with outliers increased, S/EGA and S/PA were associated with overfactoring, whereas S/MAP led to underfactoring. The PB/PA, PB/EGA, Ht/PA, and Ht/EGA methods had a mean number of factors very close to the correct value of 4 across conditions.

3.3. Empirical Example

In order to demonstrate how the methods that have been included in this study can be used in practice, demonstrations using two empirical datasets are presented. For both examples, the data consist of scores for 24 subscales from the Wechsler test of cognitive ability. Based on prior research, subscales 1–6 belong to factor 1, subscales 7–12 to factor 2, subscales 13–18 to factor 3, and subscales 19–24 to factor 4. Each subscale is normed to have a mean of 100 and a standard deviation of 15. The sample for the first example consists of 260 college undergraduates who completed the assessment, whereas the second sample consisted of 107 undergraduates who completed the assessment during a different semester from the first group. Given the results of the simulation study described above, results for EGA, PA, and MAP using the standard correlation matrix are provided, as well as correlation matrices based on the PB and Ht techniques.

Table 6 includes the mean, standard deviation, minimum, and maximum correlation values for the off-diagonal elements of the full correlation and each subscale block for the standard, percentage bend, and heavy-tailed correlation matrices.

The blocks represent the sets of subscales that theoretically belong to the same latent trait in the population (i.e., subscales 1–6, 7–12, 13–18, and 19–20). Because visualizing the full correlation matrices is very difficult, the descriptive statistics are presented here as a way of characterizing them. Perhaps the clearest trend present in these results is that, for both example datasets, the standard deviations of the PB and Ht correlations are slightly smaller than those of the standard correlation matrix. This result would be anticipated,

given that both PB and Ht are designed to remove and/or down-weight outliers. In addition, the mean correlation values for PB and Ht are either comparable to or slightly smaller (but never larger) than those of the standard correlation matrix for both example datasets. It should be noted that any such differences are not large at all, however.

The number of factors returned for the example data by each combination of extraction and outlier methods appears in Table 7.

EGA yielded the correct number (four) for each approach to dealing with the outliers (PB and Ht) for both example datasets. In contrast, PA was correct for both PB and Ht for the larger sample, but the standard correlation matrix did not yield accurate results. For the smaller sample, PA yielded the correct number of factors for Ht but not for PB. MAP consistently underestimated the number of factors to retain. These results mirror those from the simulation study, where PA and EGA yielded more accurate results than MAP, as did PB and Ht in combination with PA.

Table 8 includes the factor analysis results for each outlier method by subscale.

A cut-off of loadings greater than 0.3 was used to determine whether a subscale belonged to a factor. Specifically, the subscales to which each subscale loaded, based on a maximum likelihood factor extraction with promax rotation for four factors, are included in the table. In general, the results are quite accurate, regardless of the correlation matrix used for the analysis. The subscales that belong to the same latent trait in the population were generally also found to belong to the same population based on the factor analysis. The only exceptions to this outcome were subscale 13 for ML standard and subscale 17 for ML PB, neither of which loaded on any factor. None of the subscales loaded on more than one factor. With respect to EGA, the results were completely accurate for both the PB and Ht outlier methods. On the other hand, for the EGA standard combination, subscales 5 and 23 were not associated with the network cluster common to the rest of the items with which they were associated within the population.

4. Discussion

Prior research [3,6–8,13,36] has shown that the presence of outliers among the observed indicator variables can lead to incorrect statistical results regarding the number of factors to retain. In particular, outliers have been associated with a tendency of commonly used techniques, such as PA, to overfactor the models [6]. One approach that has been suggested for dealing with this problem is for researchers to use a robust estimate of the covariance or correlation matrix among the indicators as the basis for factor extraction [8]. Therefore, the goal of this study was to compare the performance of several methods for estimating the covariance matrix in the presence of outliers combined with multiple techniques for determining the number of factors to retain and thereby to make recommendations to practitioners faced with the problem of outlying observations.

Based on the results presented above, it appears that the PA or EGA extraction methods coupled with the percentage bend or heavy-tail outlier techniques yielded very high rates of accuracy for the number of factors to extract under the simulation conditions used in this study. These combinations had accuracy rates of 90% or higher across many of the conditions included in this study. Indeed, even when there were outliers in all of the observed variables and up to 15% of the sample was contaminated, percentage bend/PA, percentage bend/EGA, heavy-tail/PA, and heavy-tail/EGA correctly indicated the number of factors to retain in over 90% of cases. This high degree of accuracy was evident regardless of the degree of contamination in the means and standard deviations and across the sample sizes used in the study. In contrast, MAP, particularly when coupled with the standard correlation matrix and when 12 or 24 indicators were contaminated, had substantially lower accuracy rates than the other combinations of methods included in this study. With respect to the types of errors made by these approaches, the standard/PA and standard/EGA combinations led to overfactoring, whereas the standard/MAP combination led to underfactoring.

The results described here are similar to those reported in earlier research examining the impact of outliers on factor extraction. [36] also found that overfactoring was an issue when outliers were present in the data. [6] indicated that the Chi-square test was associated with overfactored solutions and that the combination of standard/PA and standard/MAP were more resistant to the presence of outliers, although not completely so. This result was also found in the current study, particularly for standard/PA, which tended to yield more accurate results than standard/MAP but was not as accurate as PA coupled with a robust covariance matrix. In addition, prior research addressing the issue of outliers in factor analysis did not investigate the performance of EGA, which has been found to be a promising tool when outliers are not present [30–32]. The current study revealed that EGA is also an effective tool in the presence of outliers, particularly when coupled with either percentage bend or heavy-tail-derived correlation matrices. Across study conditions, EGA performed as well as PA with these robust correlation matrices and better than MAP. In summary, the results of this study buttress earlier findings with respect to the use of the standard correlation/covariance matrix for factor extraction in the presence of outliers and extend it by examining the performance of EFA extraction using several robust alternatives.

Limitations and Future Research

The current study has built upon the foundation of research examining the impact of outliers on factor extraction methods that have been previously constructed [6,8,36]. As with any study, it has limitations that must be considered when interpreting the results. First, the simulation was limited to a scenario involving four latent variables with normally distributed indicator variables. Future research should examine the performance of various extraction methods for different factor structures and indicator variable distributions. Researchers often employ EFA with ordinal indicators, for example, and so more work needs to be done regarding outliers in this context. In addition, the type of outliers needs to be expanded. Second, the current work focused on outliers caused by a mixture of populations within the same sample. These mixtures were a function of differing means and/or standard deviations. However, outliers may also be caused by differences in the shape of the underlying distributions (e.g., skewness or kurtosis) for the contaminated sample or differences in the nature of relationships among the indicators. Thus, future research should examine how such differences might impact the performance of factor extraction methods coupled with robust covariance estimates. A related limitation of this study is that the approach used to induce outliers, which was borrowed from earlier work in the field [6], reflects fairly large outlier effects, particularly in regard to the mean shift of 3.0. As noted above, these values were selected because they were used in prior studies upon which the current one is built. In addition, they represent a fairly moderate (1.5) set of outliers, as well as a more extreme outlier condition (3.0), which was a goal of the current study. The results presented above revealed that some of the techniques for dealing with outliers were able to successfully ameliorate the most deleterious impact of outliers on the accuracy of methods for determining the number of factors to retain. Nonetheless, it is also true that the 3.0 mean shift condition, in particular, is quite extreme in nature. Therefore, future research should consider mean shift values between 1.5 and 3.0 in order to examine how the various outlier detection and factor number determination approaches perform in cases where outliers are clearly present in the data but are perhaps not so extreme.

Fourth, although an attempt was made to include a wide array of both factor determination methods and robust estimates of the covariance matrix, it is certainly the case that these were not exhaustive in nature. Specifically, a wider array of settings for minimum volume ellipsoid, minimum covariance determinant, or Winsorization could be used to ascertain their performance. In the current study, the defaults present in the software were used. While this seems to be a reasonable approach for a first examination of their performance when combined with the various factor determination methods, future work should expand upon them. Finally, it would be of interest to expand the outcomes to include the accuracy of factor model parameter estimates (e.g., loadings, error variances) when using the various robust

covariance matrix estimates. Early work on the impact of outliers (e.g., [8]) did study these parameter estimates with outliers and the standard covariance matrix. Thus, future work could expand upon both those earlier studies and the current one by investigating parameter estimation accuracy using robust covariance matrices.

5. Conclusions

Based on the results of this study, it would appear that researchers may be best served using a robust estimate of the covariance matrix when outliers are known to be present in the data. This recommendation is particularly salient for cases with a relatively large proportion of the sample containing outliers across a large number of the indicators being used. In addition, of the methods used here, percentage bend and heavy tail would seem to be optimal for this purpose. Either of these covariance matrices can be easily obtained using the R software package. The PA and EGA techniques for determining the number of factors to retain are also easy to employ using R, making the optimal combinations identified in this study readily available to researchers. Based on the results of this study, even if the problem of outliers turns out not to be serious, researchers using the combinations of percentage bend or heavy tail with PA or EGA will still likely obtain accurate information with respect to the number of factors to retain. Finally, it is recommended that researchers screen their data for outliers and take remedial action when necessary. Factor extraction using the standard covariance matrix when outliers are present is likely to yield inaccurate results with respect to the number of factors to retain.

Open data statement: Example data and R code for conducting the analyses described in this manuscript are available at the Open Science Foundation: <https://osf.io/kwq79/>, 1 August 2024.

Funding: This research received no external funding.

Data Availability Statement: Simulation code is available at <https://holmesfinch.substack.com/>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anderson, T.W.; Rubin, H. Statistical Inference in Factor Analysis. *Proc. Third Berkeley Symp. Math. Stat. Probab.* **1956**, *5*, 111–150.
2. Bartlett, M.S. Tests of significance in factor analysis. *Br. J. Psychol.* **1950**, *3*, 77–85. [[CrossRef](#)]
3. Bollen, K.A. Outliers and improper solutions: A confirmatory factor analysis example. *Sociol. Methods Res.* **1987**, *15*, 375–384. [[CrossRef](#)]
4. Braeken, J.; van Assen, M.A.L.M. An empirical Kaiser criterion. *Psychol. Methods* **2017**, *22*, 450–466. [[CrossRef](#)]
5. Lawley, D.N. A General Method for Approximating to the Distribution of the Likelihood Ratio Criteria. *Biometrika* **1956**, *43*, 295–303. [[CrossRef](#)]
6. Liu, Y.; Zumbo, B.D. Impact of outliers arising from unintended and unknowingly included subpopulations on the decisions about the number of factors in exploratory factor analysis. *Educ. Psychol. Meas.* **2012**, *72*, 388–414. [[CrossRef](#)]
7. Liu, Y.; Zumbo, B.D.; Wu, A.D. A demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. *Educ. Psychol. Meas.* **2012**, *72*, 181–199. [[CrossRef](#)]
8. Pison, G.; Rousseeuw, P.J.; Filzmoser, P.; Croux, C. Robust factor analysis. *J. Multivar. Anal.* **2003**, *84*, 145–172. [[CrossRef](#)]
9. Raiche G An R Package for Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test. R package Version 2.3.3.1. 2022. Available online: <https://CRAN.R-project.org/package=EFA.dimensions> (accessed on 8 July 2024).
10. Raiche, G. *R Package*, Version 0.1.7.7. EFA. Dimensions: Exploratory Factor Analysis Functions for Assessing Dimensionality. 2015.
11. Yuan, K.H.; Bentler, P.M. Improving parameter tests in covariance structure analysis. *Comput. Stat. Data Anal.* **1997**, *26*, 177–198. [[CrossRef](#)]
12. Zoski, K.W.; Jurs, S. Using Multiple Regression to Determine the Number of Factors to Retain in Factor Analysis. *Mult. Linear Regres. Viewp.* **1993**, *20*, 5–9.
13. Zygmunt, C.; Smith, M.R. Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *Quant. Methods Psychol.* **2014**, *10*, 40–55. [[CrossRef](#)]
14. Bakker, M.; Wicherts, J.M. Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE* **2014**, *9*, e103360. [[CrossRef](#)] [[PubMed](#)]
15. Jöreskog, K.G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **1969**, *34*, 183–202. [[CrossRef](#)]

16. Satorra, C.; Bentler, P.M. Corrections to test statistics and standard errors in covariance structure analysis. In *Latent Variable Analysis: Applications for Developmental Research*; von Eye, A., Clogg, C.C., Eds.; Sage: Thousand Oaks, CA, USA, 1994; pp. 399–419.
17. Browne, M.W. Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stat. Psychol.* **1984**, *37*, 62–83. [[CrossRef](#)] [[PubMed](#)]
18. Muthén, B.O.; du Toit, S.H.C.; Spisic, D. Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes. 1997. Available online: http://gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf (accessed on 8 July 2024).
19. Horn, J.L. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* **1965**, *30*, 179–185. [[CrossRef](#)] [[PubMed](#)]
20. Finch, W.H. *Exploratory Factor Analysis*; SAGE Publications: Thousand Oaks, CA, USA, 2019.
21. Auerswald, M.; Moshagen, M. How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychol. Methods* **2019**, *24*, 468–491. [[CrossRef](#)]
22. Fabrigar, L.R.; Wegener, D.T. *Exploratory Factor Analysis*; Oxford University Press: Oxford, UK, 2011.
23. Preacher, K.J.; MacCallum, R.C. Repairing Tom Swift’s Electric Factor Analysis Machine. *Underst. Stat.* **2003**, *2*, 13–43. [[CrossRef](#)]
24. Xia, Y. Determining the number of factors when population models can be closely approximated by parsimonious models. *Educ. Psychol. Meas.* **2021**, *81*, 1143–1171. [[CrossRef](#)] [[PubMed](#)]
25. Velicer, W.F. Determining the Number of Components from the Matrix of Partial Correlations. *Psychometrika* **1976**, *41*, 321–327. [[CrossRef](#)]
26. Caron, P.-O. Minimum Average Partial Correlation and Parallel Analysis: The Influence of Oblique Structures. *Commun. Stat.-Simul. Comput.* **2018**, *48*, 2110–2117. [[CrossRef](#)]
27. Garrido, L.E.; Abad, F.J.; Ponsoda, V. Performance of Velicer’s Minimum Average Partial Factor Retention Method with Categorical Variables. *Educ. Psychol. Meas.* **2011**, *71*, 551–570. [[CrossRef](#)]
28. Ruscio, J.; Roche, B. Determining the Number of Factors to Retain in an Exploratory Factor Analysis using Comparison Data of Known Factorial Structure. *Psychol. Assess.* **2012**, *24*, 282–292. [[CrossRef](#)] [[PubMed](#)]
29. Zwick, W.R.; Velicer, W.F. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychol. Bull.* **1986**, *99*, 432–442. [[CrossRef](#)]
30. Epskamp, S.; Rhemtulla, M.; Borsboom, D. Generalized network psychometrics: Combining network and latent variable models. *Psychometrika* **2017**, *82*, 904–927. [[CrossRef](#)]
31. Golino, H.F.; Epskamp, S. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE* **2017**, *12*, e0174035. [[CrossRef](#)] [[PubMed](#)]
32. Golino, H.; Shi, D.; Christensen, A.P.; Garrido, L.E.; Nieto, M.D.; Sadana, R.; Thiagarajan, J.A.; Martinez-Molina, A. Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychol. Methods* **2020**, *25*, 292–320. [[CrossRef](#)] [[PubMed](#)]
33. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441. [[CrossRef](#)] [[PubMed](#)]
34. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [[CrossRef](#)]
35. Rousseeuw, P.J. Least median of squares regress. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880. [[CrossRef](#)]
36. Yuan, K.-H.; Marshall, L.L.; Bentler, P.M. A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika* **2002**, *67*, 95–122. [[CrossRef](#)]
37. Wilcox, R.R. *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed.; Academic Press: San Diego, CA, USA, 2012.
38. Wilcox, R.R. The percentage bend correlation coefficient. *Psychometrika* **1994**, *59*, 601–616. [[CrossRef](#)]
39. Aas, K.; Haff, I.H. The generalised hyperbolic skew Student’s *t*-distribution. *J. Financ. Econom.* **2006**, *4*, 275–309.
40. Palomar, D.P.; Zhou, R.; Wang, X. *R Package*, version 0.1.4. fitHeavyTail: Mean and Covariance Matrix Estimation under heavy Tails. 2022.
41. Tong, X.; Bentler, P.M. Evaluation of a new mean scaled and moment adjusted test statistic for SEM. *Struct. Equ. Model.* **2013**, *20*, 148–156. [[CrossRef](#)] [[PubMed](#)]
42. Van Aelst, S.; Rousseeuw, P. Minimum volume ellipsoid. *WIREs Comput. Stat.* **2009**, *1*, 71–82. [[CrossRef](#)]
43. Rocke, D.M. Robustness properties of S-estimators of multivariate location and shape in high dimension. *Ann. Stat.* **1996**, *24*, 1327–1345. [[CrossRef](#)]
44. Rocke, D.M.; Woodruff, D.L. Identification of outliers in multivariate data. *J. Am. Stat. Assoc.* **1996**, *91*, 1047–1061. [[CrossRef](#)]
45. Rosseel, Y. LAVAAN: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. [[CrossRef](#)]
46. Revelle, W. *R Package*; Version 2.3.3; Psych: Procedures for Psychological, Psychometric, and Personality Research; Northwestern University: Evanston, IL, USA, 2023; Available online: <https://CRAN.R-project.org/package=psych> (accessed on 8 July 2024).
47. Hudson, G.; Alexander, P.C. EGAnet: Exploratory Graph Analysis—A Framework for Estimating the Number of Dimensions in Multivariate Data Using Network Psychometrics. *R Package* Version 2.0.5. 2024. Available online: <https://r-ega.net> (accessed on 8 July 2024).

-
48. Mair, P.; Wilcox, R. Robust Statistical Methods in R Using the WRS2 Package. *Behav. Res. Methods* **2020**, *52*, 464–488. [[CrossRef](#)]
 49. Barrera, M.; Yohai, V.; Maronna, R.; Martin, D.; Konis, K.; Croux, C.; Haesbroeck, G.; Maechler, M.; Koller, M. Robust Statistics: Theory and Methods. R Package Version 0.1.4. 2022. Available online: <https://cran.r-project.org/web/packages/RobStatTM/index.html> (accessed on 8 July 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.