

Factor Retention in Exploratory Factor Analysis With Missing Data

Educational and Psychological
Measurement
1–21

© The Author(s) 2021

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644211022031
journals.sagepub.com/home/epmDavid Goretzko¹ 

Abstract

Determining the number of factors in exploratory factor analysis is arguably the most crucial decision a researcher faces when conducting the analysis. While several simulation studies exist that compare various so-called factor retention criteria under different data conditions, little is known about the impact of missing data on this process. Hence, in this study, we evaluated the performance of different factor retention criteria—the Factor Forest, parallel analysis based on a principal component analysis as well as parallel analysis based on the common factor model and the comparison data approach—in combination with different missing data methods, namely an expectation-maximization algorithm called Amelia, predictive mean matching, and random forest imputation within the multiple imputations by chained equations (MICE) framework as well as pairwise deletion with regard to their accuracy in determining the number of factors when data are missing. Data were simulated for different sample sizes, numbers of factors, numbers of manifest variables (indicators), between-factor correlations, missing data mechanisms and proportions of missing values. In the majority of conditions and for all factor retention criteria except the comparison data approach, the missing data mechanism had little impact on the accuracy and pairwise deletion performed comparably well as the more sophisticated imputation methods. In some conditions, especially small-sample cases and when comparison data were used to determine the number of factors, random forest imputation was preferable to other missing data methods, though. Accordingly, depending on data characteristics and the selected factor retention criterion, choosing an appropriate missing data method is crucial to obtain a valid estimate of the number of factors to extract.

¹Ludwig Maximilians University Munich, Munich, Germany

Corresponding Author:

David Goretzko, Department of Psychology, Ludwig Maximilians University Munich, Leopoldstraße 13, Munich, 80802, Germany.
Email: david.goretzko@psy.lmu.de

Keywords

missing data, exploratory factor analysis, multiple imputation, factor retention

Introduction

In recent years, psychological research has increasingly focused on the issue of missing data (e.g., West, 2001). However, this cannot be said about research relying on exploratory factor analysis (EFA) where missingness is almost always neglected (Russell, 2002). One reason for this suboptimal research practice might be the limited literature on this issue. Especially, when regarding the factor retention process (determining the number of factors in EFA), there is hardly any research evaluating missing data methods. Although, there are articles (Dray & Josse, 2015; Josse & Husson, 2012) focusing on principal component analysis (PCA), the estimation of factor loadings (Lorenzo-Seva & Van Ginkel, 2016) or the proportions of explained variance (Nassiri et al., 2018), the process of determining the number of factors is mostly ignored despite its central role within the analysis.

There is an article by McNeish (2017) that dealt with the factor retention process and missing data in small-sample conditions and a simulation study by Goretzko, Heumann, and Bühner (2019) who evaluated six missing data methods in combination with parallel analysis. Both articles showed that multiple imputation seems to be favorable over pairwise or listwise deletion practices (especially the latter), but relied either on the eigenvalue-greater-one rule (Kaiser, 1960) which should not be used (Fabrigar et al., 1999; Goretzko, Pham, & Bühner, 2019) or parallel analysis (Horn, 1965) which has been shown to be inferior to more modern approaches in some data contexts (e.g., Braeken & Van Assen, 2017; Lorenzo-Seva et al., 2011; Ruscio & Roche, 2012). However, since Auerswald and Moshagen (2019) showed that no factor retention criterion is preferable under all data conditions, it seems reasonable to assume that these criteria are affected differently by missing data and that their compatibility with missing data methods also varies.

Factor Retention Criteria

Over the years, several methods to determine the number of factors in EFA—so-called factor retention criteria—have been developed. While overly simple heuristics like the Kaiser–Guttman rule (the eigenvalue-greater-one rule; Kaiser, 1960) or the scree-test (Cattell, 1966) are considered to be outdated (e.g., Fabrigar et al., 1999), more complex and often simulation-based approaches have emerged that promise a more accurate estimation of the number of latent factors. In this article, we consider parallel analysis (first implementation by Horn, 1965), comparison data (Ruscio & Roche, 2012) as well as a factor retention approach based on machine learning (Goretzko & Bühner, 2020).

Parallel Analysis. Parallel analysis is often considered a gold standard for factor retention (inter alia as it is rather robust against distributional assumptions, see Dinno, 2009). The basic idea of parallel analysis is to compare the eigenvalues of the empirical correlation matrix with eigenvalues of simulated (or resampled) data sets to determine how many empirical eigenvalues are greater than random reference eigenvalues. For this purpose, S data sets (increasing S yields more robust reference eigenvalues) are simulated and the eigenvalues of the correlation matrix are calculated. Accordingly, parallel analysis provides eigenvalue distributions based on S values for each eigenvalue. To determine the number of factors, each empirical eigenvalue is compared with the distribution of simulated reference eigenvalues and factors are retained as long as the empirical eigenvalue is greater than a specific quantile of this distribution (the initial implementation of Horn, 1965, was based on a comparison with the mean across the S simulated data sets, while often the 95% quantile is used, Revelle, 2018).

As this general idea of parallel analysis can be implemented using simulated or resampled data, the eigenvalues of the correlation matrix, or the eigenvalues of a reduced correlation matrix based on the common factor model as well as different quantiles of the reference eigenvalue distributions, the performance of these different versions of parallel analysis can vary (Lim & Jahng, 2019).

Comparison Data. The comparison data approach by Ruscio and Roche (2012) can be seen as a special case of parallel analysis using comparison data sets that reproduce the empirical correlation matrix as closely as possible based on different factor solutions instead of using random data for comparison. This method subsequently tests different factor solutions based on the RMSE between the empirical eigenvalues and the respective eigenvalues of the comparison data sets to determine whether retaining an additional factor “significantly” increases the similarity between empirical and reference eigenvalues. With this idea of simulated comparison data and a series of significance tests, comparison data unites parallel analysis and model comparisons known from structural equation modeling. Contrary to classical parallel analysis based on normal data, this method is able to take skewed item distributions into account (for further information on the data generation, see also Ruscio & Kaczetow, 2008).

Factor Retention Using Machine Learning. Recently, a new factor retention criterion relying on extensive data simulation and machine learning modeling has been developed by Goretzko and Bühner (2020). Their idea was to simulate various data sets that cover all important data conditions of an application context—varying the sample size ($N \in (200; 1000)$), the number of manifest variables (up to 80 indicators), the number of latent factors (up to eight) as well as the loading patterns and between-factor correlations (and therefore the communalities). Then the authors calculated variables that were assumed to be related to the number of underlying factors (e.g., eigenvalues, matrix norms, and inequality measures as well as more general data

characteristics such as the sample size) for each simulated data set and stored these variables together with the known number of factors of each data set as a training data set. Afterward, an *XGBoost* (Chen & Guestrin, 2016)—a tree-based machine learning model—was trained on these data to “learn” the relationship between the data characteristics and the number of underlying latent factors. Their trained *XGBoost* model (with tuned hyperparameters, see Goretzko & Bühner, 2020) outperformed all classical factor retention criteria in a simulation study with newly simulated data, but has not been evaluated with missing data yet.

Missing Data Mechanisms and Missing Data Method

The literature on missing data distinguishes between three major missing data mechanisms—missing completely at random (*MCAR*), missing at random (*MAR*), and missing not at random (*MNAR*). Little and Rubin (2002) give a more detailed introduction to these different types of missingness. In this study, we focus on the mechanisms *MCAR*, which means that missing values occur completely due to a random process and *MAR*, which means that the missingness is dependent on variables that are observed. The latter seems to be plausible in the context of EFA, since the observed variables are usually designed to be indicators for the same latent variables (Goretzko, Heumann, & Bühner, 2019).

In cases of *MCAR* or *MAR*, several imputation methods can be used to replace the missing values and to ensure valid inference. Contrary to single imputation procedures, multiple imputation methods allow for estimating the additional imputation variance (for further readings, see Little & Rubin, 2002) and are therefore preferred in most applications. Arguably the most common framework for multiple imputation is multiple imputations by chained equations (*MICE*) also known as fully conditional specification (van Buuren et al., 2006), where missing values in one variable are iteratively imputed given all other variables and their current imputed values. Within this framework, one can use various imputation models—simple regression models, tree-based methods, or specific imputation models like predictive mean matching to predict values for the missing data (e.g., Little, 1988).

In the present article, we apply the *MICE* framework with both a random forest and predictive mean matching as imputation models. Predictive mean matching is based on a (linear) regression model applied to the observed data¹ and the regression coefficients obtained from the model are taken as expected values of a multivariate normal distribution from which artificial coefficients are then randomly drawn. These “random” coefficients are used to predict the variable that should be imputed. However, contrary to a common regression imputation approach these predicted values are not taken as the imputation values but are rather compared with each other to find the most similar observations that are not missing in the empirical data set for each observation that is missing for that variable. These similar observations are then regarded as potential “donors” whose observed values are used to impute the actual missing values by selecting one of them by chance.

The random forest is a tree-based machine learning model (Breiman, 1999) that can be used as an imputation model within *MICE* as well. This model consists of several decision trees that are grown using recursive binary splitting. In this process, a number of bootstrap samples (in general, this number is around 500 and often optimized when the predictive performance should be maximized, but for imputation purposes, it can be set to 10, see Shah et al., 2014) is drawn from the empirical data set and a single tree is built on each sample using the variable of interest as the dependent variable.² This growing process stops when each terminal node only contains observations with the same value on the dependent variable or certain termination criteria are met. The resulting tree structure can be used to predict the missing values by averaging the mean value of each of the terminal nodes to which a specific observation is assigned to. In contrast to predictive mean matching which relies on a linear and therefore additive model, the possibly complex tree-structure can reflect interactions and promises to be superior when the missing data mechanism is rather complex (Goretzko, Heumann, & Bühner, 2019).

A different approach is *Amelia* (Honaker et al., 2011), an expectation-maximization (EM) algorithm that we investigate as an alternative to *MICE*. The basic idea of *Amelia* is to combine a classic EM-algorithm with bootstrapping to induce randomness necessary for multiple imputation. On each bootstrap sample, an EM-algorithm is applied to estimate the sufficient statistics for the expected values (μ) and the variance-covariance matrix (Σ) of an assumed multivariate normal distribution (an assumption that is often made for EFA modeling as well). First, μ and Σ are initially estimated based on the respective bootstrap sample. Then during each E-step, all missing values are imputed drawing from the distribution with the current values of μ and Σ and afterward, during the M-step these parameters are reestimated given the data set with the current set of imputed values. This procedure is iterated until a convergence criterion is fulfilled (Honaker & King, 2010).

Method

We wanted to extend the work of Goretzko, Heumann, and Bühner (2019) who evaluated the performance of parallel analysis in combination with *MICE* and three different imputation models (predictive mean matching, linear regression, and random forest), the *Amelia* algorithm by Honaker et al. (2011) as well as pairwise and listwise deletion varying the sample size, the number of manifest variables, the number of latent variables as well as the missing data mechanism. As they found predictive mean matching and linear regression to perform very similar and listwise deletion to be inferior in the majority of conditions, we focused on four missing data methods (the *Amelia*, *MICE* with predictive mean matching, and a random forest implementation as well as pairwise deletion as a baseline) in this study. When Goretzko, Heumann, and Bühner (2019) dealt with the impact of missing data on the factor retention process, they only evaluated one implementation of parallel analysis, but did not cross the different missing data methods with different factor retention

criteria. Accordingly, we combined the four missing data methods with two implementations of parallel analysis based on the factor model (PA-FA; as done by Goretzko, Heumann, & Bühner, 2019) and based on PCA (PA-PCA)—both using the 95% percentile of the eigenvalue distribution, the comparison data (CD) approach by Ruscio and Roche (2012), and a new machine learning approach by Goretzko and Bühner (2020) that we retrieved from the associated OSF repository (<https://osf.io/mvrau/>). For both implementations of parallel analysis, we also compared the two combination or aggregation approaches for multiple imputed data sets presented in Goretzko, Heumann, and Bühner (2019). One aggregation strategy (that we will call the *mode* approach) is based on the idea that the factor retention is done on each imputed data set and the factor solution that is proposed for the majority of imputed data sets (i.e., the mode of the distribution of the suggested numbers of factors across all imputed data sets) is used as the result for the empirical data set. The other approach (referred to as the *cor* approach) is based on an averaged correlation matrix (i.e., the correlation matrix is calculated for each imputed data set and the resulting matrices are averaged element-wise). The second strategy can also be implemented for covariance matrices as it was suggested by Nassiri et al. (2018).

Since the CD approach relies on the item distributions, averaging the correlation matrices of the different imputed data sets was not feasible (which is also the case for the machine learning approach as it relies on features that are based on the raw data as well). For this reason, we used only the *mode* approach—the most frequent factor solution across the multiple imputed data sets as the final solution for the initial data set—for both CD and the machine learning method.

Data Simulation

We slightly altered the simulation design of Goretzko, Heumann, and Bühner (2019).³ For our study, normal data were simulated for three sample sizes ($N = 250, 500, 1000$), four numbers of variables ($p = 16, 24, 36, 48$), three numbers of factors ($k = 2, 4, 6$), different values of interfactor correlations ($\rho = 0, 0.2, 0.4$), two missing data mechanisms (*MCAR*, *MAR*), and two proportions of missing values ($m = 10\%, 25\%$). In total 324 data conditions were evaluated excluding conditions with unusual variables-to-factor ratios (e.g., $k = 2$ and $p = 48$ or $k = 6$ and $p = 16$).

The data simulation was conducted with *R* (R Core Team, 2018) following the procedure of Goretzko, Heumann, and Bühner (2019). In a first step, the true factor patterns for our simulation based on standardized primary loadings between 0.5 and 0.7 and cross-loadings between 0 and 0.1 were drawn.⁴ Then a population correlation matrix Σ was calculated based on this pattern matrix and the respective between-factor correlations ($\Sigma = \Lambda\Phi\Lambda^T + \Psi^2$ with $\Psi^2 = \mathbb{1}_{p \times p} - \text{diag}(\Lambda\Phi\Lambda^T)$) which was then used to draw the data samples for given N , p and k with the *mvtnorm* package (Genz et al., 2018). We then induced missingness (*MCAR* or *MAR*) with the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011) according to Goretzko, Heumann, and Bühner (2019) who relied on Brand (1999) using the *ampute*-function

and two different sets of missingness patterns. When $p = 16$, we used two missing data patterns—either the first or the last eight variables contained missingness. When $p > 16$, $t = \lceil p \cdot m \rceil$ (which is the rounded-up product of the number of manifest variables p and the proportion of missingness m , i.e., when $p = 24$ and $m = 0.25$, $t = 6$) patterns were simulated for each condition (on average 45% of manifest variables contained missingness in each pattern). Each data set without missingness was then split into t subdata sets, so that missing values were introduced in each subset according to its own missing data pattern and the proportion of missingness set in the respective simulation condition. For further readings on this multivariate approach to induce missingness we refer to Schouten et al. (2018).

Evaluating the Missing Data Method

Missing values were either treated with pairwise deletion, imputed five times with *Amelia* or imputed five times with *MICE* in combination with predictive mean matching or random forest imputation. Then parallel analysis based on a PCA (PA-PCA) as well as parallel analysis based on the common factor model (PA-FA) using the 95% percentile of the sampled eigenvalue distribution as implemented in the *psych* package (Revelle, 2018), CD with functions provided by Ruscio and Roche (2012) and the machine learning approach by Goretzko and Bühner (2020) called *Factor Forest* (FF; see <https://osf.io/mvrau/> for the material) were applied to all imputed data sets and the most frequent solution was used as the aggregated solution for the initial data set (this approach was denoted the *mode* approach in Goretzko, Heumann, & Bühner, 2019). Averaging the correlation matrices was only possible for both implementations of the parallel analysis, so we considered 22 combinations of factor retention criteria and missing data methods (four retention criteria for pairwise deletion and six procedures for each imputation method).

For all these combinations, the suggested number of factors was averaged over all 500 replications of each data condition. We collected the accuracy and proportions of under- and overfactoring and compared all procedures (factor retention criterion + missing data method) to find the best approach for each condition.

Results

The overall accuracy of PA-FA was greater than 90% for all missing data methods (91.08%-98.79% for the *mode* approach and 98.22%-99.14% for the *cor* approach). While FF also reached 90% overall accuracy for all missing data methods (FF with predictive mean matching [*pmm*] yielded the lowest accuracy of 91.21%) and PA-PCA was able retain the correct number of factors with an overall accuracy greater than 85% for all missing data methods (*mode* and *cor* approach), CD showed a very poor performance when combined with *pmm*, the *Amelia* algorithm (*em*) or pairwise deletion (*pair*)—combinations with an overall accuracy of less than 70%. Only in combination with random forest imputation (*rf*), CD yielded a high overall accuracy

Table 1. Overall Accuracy of the Factor Retention Criteria in Combination With Different Missing Data Methods.

| Criterion | pair | pmm | rf | em |
|-------------------|-------|-------|-------|-------|
| Mode | | | | |
| FF | 0.966 | 0.912 | 0.984 | 0.959 |
| PA _{FA} | 0.983 | 0.911 | 0.988 | 0.975 |
| PA _{PCA} | 0.913 | 0.899 | 0.876 | 0.896 |
| CD | 0.656 | 0.549 | 0.962 | 0.699 |
| Cor | | | | |
| PA _{FA} | NA | 0.982 | 0.985 | 0.991 |
| PA _{PCA} | NA | 0.907 | 0.874 | 0.896 |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation, and em for the Amelia algorithm. Mode and Cor indicate which aggregation strategy was used for PA. FF = Factor Forest; PA_{FA} = parallel analysis based on common factor model; PA_{PCA} = parallel analysis based on principal component analysis; CD = comparison data; NA = not applicable.

(96.18%). The overall accuracy for each combination of factor retention criterion and missing data method is displayed in Table 1.

PA-FA and FF showed (nearly) no bias when combined with *rf*, *pair*, or *em*, but a slight tendency to overfactor when *pmm* was used (for PA-FA this was the case for the *mode* approach; there was no bias with *pmm* when the *cor* approach was used). PA-PCA underestimated the number of factors on average and showed the highest bias in combination with *rf* (independent of the aggregation strategy *mode* or *cor*). CD tended to overfactor (especially when combined with *pmm*) with all missing data methods except from *rf*. In Table 2, the estimated bias of the factor retention (the tendency of over- or underfactoring, i.e., the average deviation of the suggested number of factors from the true number of latent factors k) is presented for each combination of criterion and missing data method.

The missing data mechanism (*MCAR* vs. *MAR*) had almost no impact on the performance of the factor retention criteria. The overall accuracy of each combination of factor retention criterion and missing data method differed less than one percentage point between *MCAR* and *MAR* (second highest difference 0.01) with one exception, CD with *pair* reaching a 4.83 percentage points higher accuracy with *MCAR* than with *MAR*.

As expected, the higher the sample size N was, the more accurate all combinations of factor retention criteria and missing data methods were. With $N = 1000$, FF and PA-FA (*mode* and *cor* approach) yielded (nearly) perfect accuracy for all missing data methods, while PA-PCA reached approximately 95% accuracy independently of the aggregation strategy (*mode* or *cor*) and the missing data method. The performance of CD highly varied between an accuracy of 70.22% with *pmm* and 98.85% with *rf*. With $N = 250$, all methods yielded considerably lower accuracies. In Table 3, the overall accuracy of all combinations of factor retention criteria and missing data

Table 2. Estimated Bias of the Factor Retention Criteria in Combination With Different Missing Data Methods.

| Criterion | pair | pmm | rf | em |
|-------------|--------|--------|--------|--------|
| Mode | | | | |
| FF | 0.070 | 0.193 | 0.018 | 0.076 |
| PA_{FA} | 0.013 | 0.106 | -0.015 | 0.018 |
| PA_{PCA} | -0.178 | -0.175 | -0.296 | -0.223 |
| CD | 0.511 | 0.904 | 0.007 | 0.430 |
| Cor | | | | |
| PA_{FA} | NA | 0.013 | -0.021 | -0.001 |
| PA_{PCA} | NA | -0.195 | -0.311 | -0.231 |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation, and em for the Amelia algorithm. Mode and Cor indicate which aggregation strategy was used for PA. FF = Factor Forest; PA_{FA} = parallel analysis based on common factor model; PA_{PCA} = parallel analysis based on principal component analysis; CD = comparison data; NA = not applicable.

Table 3. Overall Accuracy of the Factor Retention Criteria in Combination With Different Missing Data Methods for Small Sample Sizes ($N = 250$).

| Criterion | pair | pmm | rf | em |
|-------------|-------|-------|-------|-------|
| Mode | | | | |
| FF | 0.908 | 0.772 | 0.953 | 0.886 |
| PA_{FA} | 0.959 | 0.777 | 0.965 | 0.930 |
| PA_{PCA} | 0.846 | 0.814 | 0.784 | 0.822 |
| CD | 0.546 | 0.422 | 0.919 | 0.547 |
| Cor | | | | |
| PA_{FA} | NA | 0.954 | 0.956 | 0.975 |
| PA_{PCA} | NA | 0.836 | 0.777 | 0.823 |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation and em for the Amelia algorithm. Mode and Cor indicate which aggregation strategy was used for PA. FF = Factor Forest; PA_{FA} = parallel analysis based on common factor model; PA_{PCA} = parallel analysis based on principal component analysis; CD = comparison data; NA = not applicable.

methods is displayed for these small-sample conditions with $N = 250$. In small-sample-size conditions, *pmm* yielded the lowest accuracy. While the combination of *pmm* and PA_{FA} -*cor* reached an accuracy of 95.40%, FF, PA_{FA} -*mode* as well as PA_{PCA} (both aggregation strategies) yielded notably lower accuracies (between 77.23% and 83.65%) and CD was accurate only 42.19% of the time. *rf* provided very good results when combined with FF, PA_{FA} and CD, whereas in combination with PA_{PCA} all other missing data methods showed (slightly) better results in these conditions.

Interfactor correlations generally worsened the factor retention process—the higher these correlations were, the lower the accuracy was (see Table 4). This was striking in the case of PA-PCA in combination with *rf* (both *cor* and *mode*) which reached a very high accuracy in conditions with no or little between-factor correlations ($\rho=0$: $Acc_{PA-PCA-cor}=99.55\%$ and $Acc_{PA-PCA-mode}=99.66\%$; $\rho=0.2$: $Acc_{PA-PCA-cor}=93.83\%$ and $Acc_{PA-PCA-mode}=94.09\%$), but performed way worse when substantial between-factor correlations were present ($\rho=0.4$: $Acc_{PA-PCA-cor}=68.90\%$ and $Acc_{PA-PCA-mode}=69.15\%$). Contrary, CD benefited from correlated factors as its performance improved for three of four missing data methods (CD and *rf* reached an accuracy of approximately 95% independently of the interfactor correlation). Although, CD in combination with *pair* ($\rho=0$: $Acc_{CD}=63.39\%$ vs. $\rho=0.4$: $Acc_{CD}=67.26\%$), *em* ($\rho=0$: $Acc_{CD}=63.25\%$ vs. $\rho=0.4$: $Acc_{CD}=76.04\%$), and *pmm* ($\rho=0$: $Acc_{CD}=49.50\%$ vs. $\rho=0.4$: $Acc_{CD}=60.20\%$) provided better results with correlated factors than with orthogonal factors, their performance was inferior to all other combinations of factor retention criteria and missing data methods.

Figure 1 displays the overall accuracy of the factor retention process for each combination of criterion (and aggregation level) and missing data method. When 10% of the data was missing, the number of factors could be determined quite accurately by all methods. FF and PA-FA (*cor* and *mode* approach) yielded almost perfect accuracy, while CD in combination with *pmm*, *pair*, or *em* performed notably worse (the respective conditions are removed from the plot as the CD reached an accuracy of less than 50% for all values of k ; CD in combination with *rf* showed a comparably high accuracy though). PA-PCA (both *cor* and *mode*) was able to retain two and four factors with nearly perfect accuracy, but was inferior to all other factor retention criteria when $k=6$ (independently of the missing data method).

In conditions with 25% missingness, CD had very poor accuracies in combination with *em*, *pair*, and *pmm*, but was competitive when combined with *rf* (even though CD + *rf* was slightly inferior to PA-FA and FF with an error-rate of approximately 10% in conditions with six factors). FF struggled to correctly identify the number of factors in conditions with $k=2$ and 25% missingness. However when combined with *rf* instead of *em*, *pmm*, or *pair*, FF reached an accuracy of 92.82%, while all other methods performed even better. PA-PCA showed the same pattern as in conditions with 10% missingness—independently of the missing data method, it reached a high accuracy when the true number of factors was rather low ($k \in [2, 4]$) and yielded a substantially smaller accuracy when $k=6$.

Choosing the Best Combination of Factor Retention Criterion and Missing Data Method

For readers who are interested in choosing a combination of factor retention criterion and missing data method that performs best for a specific data context, we provide a more detailed presentation of our results in Tables 5 and 6. There, the accuracy of

Table 4. Accuracy of All Combinations of Factor Retention Criteria and Missing Data Methods for Interfactor Correlations ρ .

| Criterion | $\rho = 0.0$ | | | $\rho = 0.2$ | | | $\rho = 0.4$ | | | |
|------------------------|--------------|-------|-------|--------------|-------|-------|--------------|-------|-------|-------|
| | pair | pmm | rf | pair | pmm | rf | pair | pmm | rf | em |
| | | | em | | | em | | | em | |
| Mode | | | | | | | | | | |
| FF | 0.977 | 0.925 | 0.996 | 0.972 | 0.919 | 0.991 | 0.947 | 0.893 | 0.966 | 0.947 |
| PA _{FA-mode} | 0.987 | 0.902 | 0.999 | 0.987 | 0.911 | 0.996 | 0.975 | 0.919 | 0.969 | 0.976 |
| PA _{PCA-mode} | 0.998 | 0.974 | 0.997 | 0.970 | 0.965 | 0.941 | 0.770 | 0.759 | 0.692 | 0.732 |
| CD | 0.634 | 0.495 | 0.969 | 0.661 | 0.550 | 0.969 | 0.673 | 0.602 | 0.948 | 0.760 |
| Cor | | | | | | | | | | |
| PA _{FA-cor} | NA | 0.982 | 0.999 | NA | 0.985 | 0.994 | NA | 0.979 | 0.962 | 0.984 |
| PA _{PCA-cor} | NA | 0.998 | 0.996 | NA | 0.967 | 0.938 | NA | 0.757 | 0.689 | 0.732 |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation and em for the Amelia algorithm. Mode and Cor indicate which aggregation strategy was used for PA. FF = Factor Forest; PA_{FA} = parallel analysis based on common factor model; PA_{PCA} = parallel analysis based on principal component analysis; CD = comparison data; NA = not applicable.

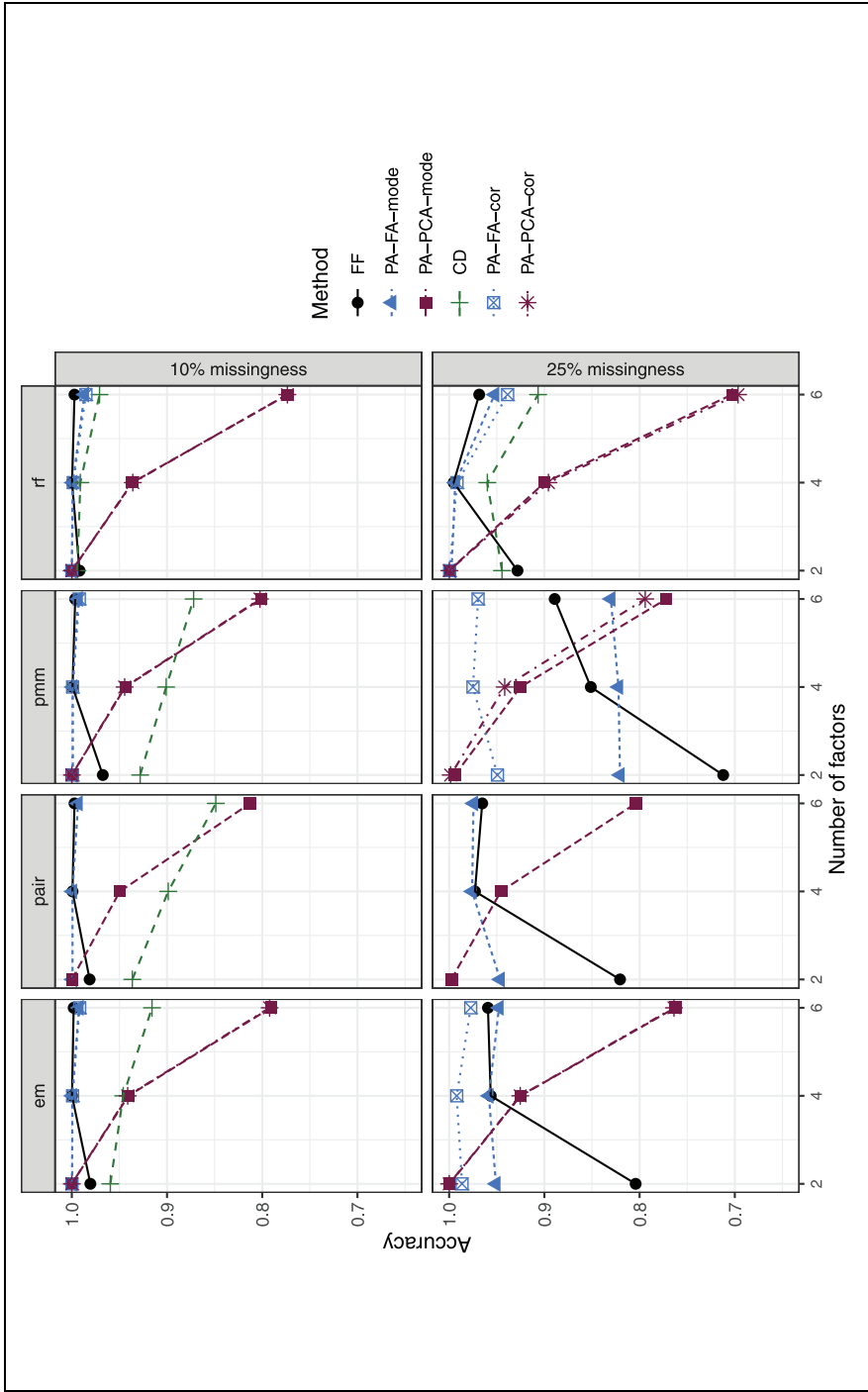


Figure 1. Accuracy of all combinations of factor retention criteria and missing data methods for different amounts of missingness (10% vs. 25%) and different factor solutions (CD + em/pmm/pair with 25% missingness as the accuracy was less than 50%).

Table 5. Accuracy of All Combinations of Factor Retention Criteria and Missing Data Methods for Different Sample Sizes, Numbers of Manifest Variables and Proportions of Missingness (Part 1).

| N | p | $\%_{\text{miss}}$ | FF | | | | | PA-FA-mode | | | | | PA-PCA-mode | | | | | CD | | | | |
|-------|----|--------------------|------|------|------|------|------|------------|------|------|------|------|-------------|------|------|------|------|------|------|------|------|----|
| | | | em | pair | pmm | rf | em | pair | pmm | rf | em | pair | pmm | rf | em | pair | pmm | rf | em | pair | pmm | rf |
| 250 | 16 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.83 | 0.84 | 0.83 | 0.82 | 0.82 | 0.91 | 0.87 | 0.85 | 0.97 | |
| 250 | 16 | 25 | 0.92 | 0.91 | 0.81 | 0.98 | 0.90 | 0.88 | 0.72 | 0.96 | 0.81 | 0.83 | 0.82 | 0.77 | 0.31 | 0.30 | 0.10 | 0.88 | 0.88 | 0.10 | 0.88 | |
| 250 | 24 | 10 | 0.96 | 0.95 | 0.93 | 0.98 | 0.98 | 0.98 | 0.98 | 0.96 | 0.78 | 0.80 | 0.79 | 0.76 | 0.87 | 0.83 | 0.81 | 0.94 | 0.94 | 0.81 | 0.94 | |
| 250 | 24 | 25 | 0.58 | 0.63 | 0.45 | 0.77 | 0.87 | 0.91 | 0.65 | 0.90 | 0.76 | 0.79 | 0.78 | 0.69 | 0.23 | 0.26 | 0.06 | 0.82 | 0.82 | 0.06 | 0.82 | |
| 250 | 36 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.83 | 0.84 | 0.84 | 0.82 | 0.85 | 0.78 | 0.75 | 0.97 | 0.97 | 0.75 | 0.97 | |
| 250 | 36 | 25 | 0.93 | 0.96 | 0.70 | 0.99 | 0.87 | 0.95 | 0.50 | 0.95 | 0.82 | 0.84 | 0.77 | 0.76 | 0.18 | 0.24 | 0.03 | 0.89 | 0.89 | 0.03 | 0.89 | |
| 250 | 48 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.92 | 0.94 | 0.93 | 0.88 | 0.85 | 0.79 | 0.73 | 0.98 | 0.98 | 0.73 | 0.98 | |
| 250 | 48 | 25 | 0.82 | 0.93 | 0.39 | 0.99 | 0.84 | 0.96 | 0.33 | 0.99 | 0.89 | 0.93 | 0.79 | 0.83 | 0.17 | 0.30 | 0.03 | 0.94 | 0.94 | 0.03 | 0.94 | |
| 500 | 16 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.89 | 0.94 | 0.91 | 0.91 | 0.99 | 0.91 | 0.91 | 0.99 | |
| 500 | 16 | 25 | 1.00 | 0.99 | 0.98 | 1.00 | 0.98 | 0.96 | 0.91 | 1.00 | 0.87 | 0.91 | 0.90 | 0.85 | 0.43 | 0.37 | 0.14 | 0.92 | 0.92 | 0.14 | 0.92 | |
| 500 | 24 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.88 | 0.88 | 0.88 | 0.88 | 0.95 | 0.89 | 0.91 | 0.99 | 0.91 | 0.91 | 0.99 | |
| 500 | 24 | 25 | 0.95 | 0.94 | 0.83 | 1.00 | 0.99 | 0.99 | 0.93 | 1.00 | 0.86 | 0.88 | 0.88 | 0.84 | 0.36 | 0.33 | 0.09 | 0.96 | 0.96 | 0.09 | 0.96 | |
| 500 | 36 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.94 | 0.96 | 0.95 | 0.92 | 0.96 | 0.89 | 0.92 | 1.00 | 0.92 | 0.92 | 1.00 | |
| 500 | 36 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 0.95 | 0.95 | 0.85 | 0.41 | 0.39 | 0.11 | 0.98 | 0.98 | 0.11 | 0.98 | |
| 500 | 48 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.92 | 0.95 | 1.00 | 0.92 | 0.95 | 1.00 | |
| 500 | 48 | 25 | 1.00 | 1.00 | 0.96 | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.50 | 0.51 | 0.16 | 0.99 | 0.99 | 0.16 | 0.99 | |
| 1,000 | 16 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.94 | 0.96 | 0.99 | 0.94 | 0.96 | 0.99 | |
| 1,000 | 16 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 0.94 | 0.97 | 0.96 | 0.94 | 0.61 | 0.47 | 0.28 | 0.93 | 0.93 | 0.28 | 0.93 | |
| 1,000 | 24 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.89 | 0.90 | 0.90 | 0.89 | 0.99 | 0.94 | 0.98 | 1.00 | 0.94 | 0.98 | 1.00 | |
| 1,000 | 24 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.89 | 0.90 | 0.89 | 0.89 | 0.65 | 0.51 | 0.29 | 0.99 | 0.99 | 0.29 | 0.99 | |
| 1,000 | 36 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 | 0.99 | 1.00 | 0.96 | 0.99 | 1.00 | |
| 1,000 | 36 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.82 | 0.67 | 0.50 | 1.00 | 0.82 | 0.50 | 1.00 | |
| 1,000 | 48 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | |
| 1,000 | 48 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.92 | 0.79 | 0.69 | 1.00 | 0.92 | 0.69 | 1.00 | |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation and em for the Amelia algorithm. PA-FA = Parallel analysis based on the factor model; PA-PCA = parallel analysis based on principal component analysis; FF = Factor Forest.

Table 6. Accuracy of All Combinations of Factor Retention Criteria and Missing Data Methods for Different Sample Sizes, Numbers of Manifest Variables and Proportions of Missingness (Part 2).

| N | p | % _{miss} | PA-FA-cor | | | PA-PCA-cor | | |
|-------|----|-------------------|-----------|------|------|------------|------|------|
| | | | em | pmm | rf | em | pmm | rf |
| 250 | 16 | 10 | 0.99 | 0.99 | 0.99 | 0.83 | 0.83 | 0.82 |
| 250 | 16 | 25 | 0.95 | 0.89 | 0.96 | 0.81 | 0.82 | 0.77 |
| 250 | 24 | 10 | 0.98 | 0.98 | 0.96 | 0.78 | 0.79 | 0.76 |
| 250 | 24 | 25 | 0.93 | 0.90 | 0.88 | 0.75 | 0.78 | 0.68 |
| 250 | 36 | 10 | 1.00 | 1.00 | 0.99 | 0.83 | 0.84 | 0.82 |
| 250 | 36 | 25 | 0.98 | 0.94 | 0.93 | 0.82 | 0.83 | 0.73 |
| 250 | 48 | 10 | 1.00 | 1.00 | 1.00 | 0.91 | 0.92 | 0.88 |
| 250 | 48 | 25 | 0.99 | 0.95 | 0.97 | 0.90 | 0.92 | 0.82 |
| 500 | 16 | 10 | 1.00 | 1.00 | 1.00 | 0.90 | 0.90 | 0.89 |
| 500 | 16 | 25 | 0.99 | 0.97 | 1.00 | 0.87 | 0.90 | 0.85 |
| 500 | 24 | 10 | 1.00 | 1.00 | 1.00 | 0.88 | 0.88 | 0.88 |
| 500 | 24 | 25 | 1.00 | 0.99 | 1.00 | 0.86 | 0.88 | 0.84 |
| 500 | 36 | 10 | 1.00 | 1.00 | 1.00 | 0.94 | 0.95 | 0.92 |
| 500 | 36 | 25 | 1.00 | 1.00 | 1.00 | 0.90 | 0.94 | 0.85 |
| 500 | 48 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 48 | 25 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.97 |
| 1,000 | 16 | 10 | 1.00 | 1.00 | 1.00 | 0.98 | 0.98 | 0.97 |
| 1,000 | 16 | 25 | 1.00 | 0.99 | 1.00 | 0.94 | 0.97 | 0.94 |
| 1,000 | 24 | 10 | 1.00 | 1.00 | 1.00 | 0.89 | 0.90 | 0.89 |
| 1,000 | 24 | 25 | 1.00 | 1.00 | 1.00 | 0.89 | 0.89 | 0.89 |
| 1,000 | 36 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1,000 | 36 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 |
| 1,000 | 48 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1,000 | 48 | 25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Note. pair stands for pairwise deletion, pmm for predictive mean matching, rf for random forest imputation and em for the Amelia algorithm. PA-FA = Parallel analysis based on the factor model; PA-PCA = parallel analysis based on principal component analysis; FF = Factor Forest.

each of the 22 combinations (factor retention criterion + missing data method + aggregation strategy) is displayed for all evaluated sample sizes, numbers of manifest variables, and different proportions of missing data. We aggregated the results for both missing data mechanisms, different levels of between-factor correlation as well as the three numbers of latent factors as these variables are usually unknown for empirical data. One can see that for many conditions, all combinations yield high accuracies (e.g., $N = 1000$, $p = 48$, and 10% missing values), while in other conditions, some combinations perform quite weakly and others are nearly perfectly accurate (e.g., when $N = 250$, $p = 48$, and the proportion of missingness equals 25%, FF should be combined with random forest imputation [99% accuracy] rather than with predictive mean matching [39% accuracy], whereas random forest imputation in

Table 7. Accuracy of the Factor Retention Criteria Different Sample Sizes and Numbers of Manifest Variables When No Data Are Missing.

| <i>N</i> | <i>p</i> | FF | PA-FA | PA-PCA | CD |
|----------|----------|------|-------|--------|------|
| 250 | 16 | 1.00 | 1.00 | 0.84 | 0.96 |
| 250 | 24 | 0.99 | 0.98 | 0.80 | 0.92 |
| 250 | 36 | 1.00 | 1.00 | 0.83 | 0.90 |
| 250 | 48 | 1.00 | 1.00 | 0.92 | 0.89 |
| 500 | 16 | 1.00 | 1.00 | 0.91 | 0.98 |
| 500 | 24 | 1.00 | 1.00 | 0.88 | 0.96 |
| 500 | 36 | 1.00 | 1.00 | 0.95 | 0.96 |
| 500 | 48 | 1.00 | 1.00 | 1.00 | 0.97 |
| 1,000 | 16 | 1.00 | 1.00 | 0.98 | 0.98 |
| 1,000 | 24 | 1.00 | 1.00 | 0.90 | 0.98 |
| 1,000 | 36 | 1.00 | 1.00 | 1.00 | 0.98 |
| 1,000 | 48 | 1.00 | 1.00 | 1.00 | 0.99 |

Note. PA-FA = parallel analysis based on the factor model; PA-PCA = parallel analysis based on principal component analysis; CD = comparison data; FF = Factor Forest.

combination with PA-PCA and the *mode*-approach was outperformed by all other missing data methods).

Baseline Comparison

For an easier interpretation of these results, the accuracy of the four factor retention criteria on comparable data without any missing values is displayed in Table 7. While FF and PA-FA estimated the number of factors correctly in almost every data set, CD had slightly lower accuracies across all conditions (mostly above 95% accuracy), while PA-PCA showed rather poor performance in small-sample conditions (see also Table 3).

All in all, most combinations provided similar results compared with the baseline performance of the factor retention criterion on fully observed data. CD in combination with the Amelia algorithm, predictive mean matching or pairwise-deletion, though, showed a more than 25 percentage points lower accuracy on average than CD's baseline performance on data without missing values (CD and predictive mean matching had a 40.70 percentage points lower accuracy), while this performance gap was in the single digits for all other combinations (CD and random forest imputation showed basically the same accuracy [$\sim 96\%$] as CD in the baseline conditions).

Discussion

In this study, 22 combinations of missing data methods (*pair*, *rf*, *em*, and *pmm*) and factor retention criteria (FF, PA-FA-*cor*, PA-FA-*mode*, PA-PCA-*cor*, PA-PCA-*mode*, and CD) were evaluated with regard to their accuracy and bias (over- or

underfactoring) in various simulated data conditions. While the choice of the missing data method had very little impact on the accuracy when PA-based factor retention or the new FF approach were used, it was crucial for an accurate factor retention when CD was used. Contrary to the findings of Goretzko, Heumann, and Bühner (2019), the missing data methods performed quite similar which can be explained by the altered simulation conditions—in this study higher primary loadings (associated with higher communalities and more reliable indicators) were used as we wanted to focus on conditions in which all factor retention criteria are able to retain the true number of factors with (nearly) perfect accuracy when no data are missing. Accordingly, the overall accuracy of each method is notably higher in this study.

McNeish (2017) and Goretzko, Heumann, and Bühner (2019) found pairwise deletion to be inferior to multiple imputation (especially random forest imputation and the Amelia algorithm in combination with the *cor* aggregation strategy)—a result that we did not find for parallel analysis in the investigated data conditions. An explanation for the relatively good performance of pairwise deletion might be the comparably “easy” conditions with (almost) perfect simple structure in this study, whereas Goretzko, Heumann, and Bühner (2019) investigated more difficult conditions with smaller primary and higher cross-loadings. In addition, McNeish (2017) focused on small-sample conditions and found pairwise deletion to be inferior to multiple imputation mainly in conditions with 60 or 120 observations, while factor retention with pairwise deletion worked similarly well when $N = 240$ (which is close to our small-sample condition).

Contrary to parallel analysis, the Factor Forest performed better with random forest imputation than with pairwise deletion. This tendency was especially recognizable in small-sample conditions ($N = 250$) that seem to be the most important for comparison to real EFA applications in psychological research (Fabrigar et al., 1999; Goretzko, Pham, & Bühner, 2019). When CD was used for factor retention, though, pairwise deletion (as well as Amelia and predictive mean matching) clearly performed worse than random forest imputation. In conditions with small sample sizes and/or high proportions of missingness (25% missing values), CD could not be seen as a valid method unless combined with random forest imputation. One explanation why CD was the only factor retention criterion that was strongly influenced by missing data could be that contrary to PA, for example, the item distributions are taken into account when simulating the comparison data sets (see Ruscio & Roche, 2012). Accordingly, in conditions with high proportions of missingness (here 25% missing values), these item distributions might be distorted if the missing values are not properly imputed.

Hence, as discussed by Goretzko, Heumann, and Bühner (2019) random forest imputation seem to be the most promising way to deal with missingness under *MAR* or *MCAR* assumption in the context of EFA (and factor retention). Nevertheless, as it showed a comparably poor performance when combined with PA-PCA (*cor* and *mode* strategy) when the number of factors got higher (here $k = 6$) and/or in conditions with substantially correlated factors (here between-factor correlation of

$\rho = 0.4$)—results that are in line with those of Goretzko, Heumann, and Bühner (2019), *rf* should not be used without considering the special application context.

When researchers want to use parallel analysis and an imputation method, they should rather use the *cor* aggregation strategy or the similar approach by Nassiri et al. (2018) instead of the *mode* approach, even though the performance differences between these two were rather small in this study. PA-FA-*cor* showed higher overall accuracies than PA-FA-*mode*, while being less biased—a tendency that yielded substantial performance differences in small-sample conditions (which again might be the most important for current psychological research practice).

Since this study is the first to investigate the interplay of different factor retention criteria and missing data methods, its focus on rather desirable data conditions with clear factor patterns and (practically) simple structure, in which an accurate factor retention is comparably easy, can be seen critically. As Goretzko, Heumann, and Bühner (2019) showed that the performance of different missing data methods differ more strongly under less favorable conditions, further research may expand the scope of this simulation study by adding other data conditions with higher cross-loadings, nonnormal data, or minor factors in the data-generating models. Another potential limitation of the current study are the proportions of missingness that were under investigation. In other simulation studies with regard to missing data much higher proportions of missing values are considered (e.g., Jochen et al., 2013), but since EFA is mostly applied when developing a questionnaire or psychological test, we would argue that the rate of item nonresponse in single questionnaires is arguably lower than in extensive surveys or settings where the questionnaire is presented after a time-consuming experiment (i.e., when solely the items have to be answered that are then used for the EFA) and therefore rarely higher than 25%. Besides, other studies on this topic used similar proportions of missingness—McNeish (2017) used up to 25% missing values, Nassiri et al. (2018) up to 30%, Lorenzo-Seva and Van Ginkel (2016) up to 15%, and Josse et al. (2011) up to 30% as well. Nevertheless, as modern instruments of data collection (e.g., mobile sensing, Schoedel et al., 2020) can yield higher proportions of missing values, further research should evaluate the influence of substantially higher missingness rates.

Conclusion

The present study evaluated different combinations of missing data methods and factor retention criteria with regard to their accuracy and potential biases (namely under- and overfactoring). For data conditions in which all compared factor retention methods are able to determine the number of factors accurately when no data are missing, all investigated missing data methods performed comparably well in combination with parallel analysis (for both tested aggregation strategies) or the factor forest. Accordingly, pairwise deletion yielded similar results as multiple imputation models based on an EM algorithm and *MICE*. However, when the comparison data approach was used for factor retention, pairwise deletion performed poorly and solely

random forest imputation within the *MICE* framework provided accurate estimates of the dimensionality. Consequently, this study shows that depending on which factor retention criterion is used to assess the dimensionality in EFA, different missing data methods may be favorable and researchers should be careful when relying on default settings such as pairwise deletion. Combining the results of this study with those of other studies, researchers are advised to compare different missing data mechanisms (to evaluate the robustness of their solution) and factor retention criteria to obtain a robust and accurate estimate of the number of factors.


Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

David Goretzko  <https://orcid.org/0000-0002-2730-6347>

Notes

1. To be more precise, all observations which are not missing for the variable that should be imputed are used—which means that all other variables in the data set that are used as predictor variables in the imputation model have to be fully observed or imputed before. For this reason, the idea of *MICE* is to sequentially impute all variables and iterate this process to get better results.
2. Each split is done using the independent variable (out of a subset of randomly drawn variables—a measure against overfitting) that best separates the two resulting subgroups with regard to the variable of interest. In other words, a split maximizes between-group differences and minimizes within-group differences considering that particular variable.
3. In this study, we used the same sample sizes and numbers of latent variables that were investigated by Goretzko, Heumann, and Bühner (2019). We further added conditions with 36 manifest variables and systematically varied the between-factor correlations, while Goretzko, Heumann, and Bühner (2019) only compared orthogonal conditions and oblique conditions with fixed between-factor correlations. Contrary to their study, we investigated two different proportions of missingness (10% and 25% vs. just 25%) and two missing data mechanisms (they evaluated four mechanisms including three different types of *MAR* that did not yield substantially different results).
4. We used higher primary and smaller cross-loadings compared with Goretzko, Heumann, and Bühner (2019) since we wanted to ensure that all evaluated factor retention criteria show nearly perfect accuracy when no data are missing. When no data were missing, FF had an overall accuracy across all conditions in this study of 99.88%, while PA-FA reached

99.76%, PA-PCA 91.16%, and CD was able to correctly identify the number of factors in 95.58% the cases.

References

- Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods, 24*(4), 468-491. <https://doi.org/10.1037/met0000200>
- Braeken, J., & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods, 22*(3), 450-466.
- Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets* [Unpublished doctoral dissertation]. Erasmus University Rotterdam, Netherlands. <https://core.ac.uk/download/pdf/18508128.pdf>
- Breiman, L. (1999, September). *Random forest*. https://machinelearning202.pbworks.com/w/file/60606349/breiman_randomforests.pdf
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research, 44*(3), 362-388. <https://doi.org/10.1080/00273170902938969>
- Dray, S., & Josse, J. (2015). Principal component analysis with missing values: A comparative survey of methods. *Plant Ecology, 216*(5), 657-667. <https://doi.org/10.1007/s11258-014-0406-z>
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2018). *mvtnorm: Multivariate normal and t distributions*. <https://CRAN.R-project.org/package=mvtnorm>
- Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods, 25*(6), 776-786. <https://doi.org/10.1037/met0000262>
- Goretzko, D., Heumann, C., & Bühner, M. (2019). Investigating parallel analysis in the context of missing data: A simulation study comparing six missing data methods. *Educational and Psychological Measurement, 80*(4), 756-774. <https://doi.org/10.1177/0013164419893413>
- Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*. Advance online publication. <https://doi.org/10.1007/s12144-019-00300-2>
- Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561-581. <https://doi.org/10.1111/j.1540-5907.2010.00447.x>
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1-47. <https://www.jstatsoft.org/article/view/v045i07>

- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185. <https://doi.org/10.1007/BF02289447>
- Jochen, H., Max, H., Tamara, B., & Wilfried, L. (2013). Multiple imputation of missing data: A simulation study on a binary response. *Open Journal of Statistics*, 3(5), 370-378. <https://doi.org/10.4236/ojs.2013.35043>
- Josse, J., & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6), 1869-1879. <https://doi.org/10.1016/j.csda.2011.11.012>
- Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3), 231-246. <https://doi.org/10.1007/s11634-011-0086-7>
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151. <https://doi.org/10.1177/001316446002000116>
- Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, 24(4), 452-467. <https://doi.org/10.1037/met0000230>
- Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287-296. <https://doi.org/10.2307/1391878>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.
- Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, 46(2), 340-364. <https://doi.org/10.1080/00273171.2011.564527>
- Lorenzo-Seva, U., & Van Ginkel, J. R. (2016). Multiple imputation of missing values in exploratory factor analysis of multidimensional scales: estimating latent trait scores. *Anales de Psicología/Annals of Psychology*, 32(2), 596-608. <https://doi.org/10.6018/analesps.32.2.215161>
- McNeish, D. (2017). Exploratory factor analysis with small samples and missing data. *Journal of Personality Assessment*, 99(6), 637-652. <https://doi.org/10.1080/00223891.2016.1252382>
- Nassiri, V., Lovik, A., Molenberghs, G., & Verbeke, G. (2018). On using multiple imputation for exploratory factor analysis of incomplete data. *Behavior Research Methods*, 50(2), 501-517. <https://doi.org/10.3758/s13428-017-1013-4>
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2018). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Ruscio, J., & Kacetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, 43(3), 355-381. <https://doi.org/10.1080/00273170802285693>
- Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, 24(2), 282-292. <https://doi.org/10.1037/a0025697>
- Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin*. *Personality and Social Psychology Bulletin*, 28(12), 1629-1646. <https://doi.org/10.1177/014616702237645>
- Schoedel, R., Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., & Stachl, C. (2020). To challenge the morning lark and the night owl: Using smartphone sensing data to

- investigate day–night behaviour patterns. *European Journal of Personality*, 34(5), 733-752. <https://doi.org/10.1002/per.2258>
- Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15), 2909-2930. <https://doi.org/10.1080/00949655.2018.1491577>
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology*, 179(6), 764-774. <https://doi.org/10.1093/aje/kwt312>
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064. <https://doi.org/10.1080/10629360600810434>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67. <https://www.jstatsoft.org/v45/i03/>
- West, S. G. (2001). New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods*, 6(4), 315-316. <https://doi.org/10.1037/1082-989X.6.4.315>