Behavior Research Methods (2018) 50:501–517
https://doi.org/10.3758/s13428-017-1013-4

On using multiple imputation for exploratory factor analysis of incomplete data

Vahid Nassiri[1] · **Anikó Lovik**[1] · **Geert Molenberghs**[1,2] · **Geert Verbeke**[1,2]

Published online: 1 February 2018

Abstract

A simple multiple imputation-based method is proposed to deal with missing data in exploratory factor analysis. Confidence intervals are obtained for the proportion of explained variance. Simulations and real data analysis are used to investigate and illustrate the use and performance of our proposal.

Keywords Missing data · Multiple imputation · Exploratory factor analysis · Principal component analysis

Introduction

Exploratory factor analysis (EFA) and principal component analysis (PCA) are techniques mainly based on singular value decomposition of covariance matrices of multivariate data, e.g., a questionnaire, several measurements on a subject, etc. In multivariate data, it is very well possible that for some of the subjects, one or more of the variables are missing. One approach to deal with this phenomenon is listwise deletion, i.e., removing all subjects with missing values and only using the observed part of the data. This would lead to loss of information, and worse than that, biased estimates and conclusions when the amount of missing data is large or the data are not missing completely at random (MCAR) (Rubin, 1976). Especially when the amount of missing data is large, one may be faced with a singularity problem after removing the incomplete subjects, i.e., the number of items could become larger than the number of fully captured observations. As a result, the estimated variance-covariance matrices may turn out to be non-positive-definite.

Another approach to deal with estimating the covariance matrix of incomplete data is pairwise deletion, i.e., to use completely observed pairs. The same drawbacks as in the listwise deletion case can be considered here. Full information maximum likelihood (FIML) is another established approach to deal with incomplete data. FIML tries to maximally use the information from all subjects by allowing subject-specific parameter dimensions (Enders & Bandalos, 2001). The singularity problem discussed earlier could cause difficulties for FIML. Considering the fact that EFA is usually used in early stages of data collection, the available sample size is small. As McNeish (2016) mentioned, according to Russell (2002), 39% of studies use a sample size of less than 100 and this is between 100 and 200 for 23% of them. Adding missing data to such small samples would face methods like listwise deletion, pairwise deletion, and FIML with difficulties. This can be seen in the simulation study we have performed in this paper as well. One also can use the EM algorithm (Dempster, Laird, & Rubin 1977) to iteratively estimate the covariance matrix of the incomplete data.

Wold and Lyttkens (1969) proposed the nonlinear iterative partial least squares estimation (NIPALS) procedure, which uses an alternating weighted least-squares algorithm and estimates principal components one by one in a sequential manner. Gabriel and Zamir (1979) extended NIPALS to directly estimate the desired subspace, rather than sequentially. Iterative principal component analysis (Kiers, 1997) is used to estimate the missing values and the principal components simultaneously. The main difference between iterative PCA and NIPALS is that NIPALS tries to estimate principal components regardless of the missing values, while iterative PCA produces a single imputation for the missing values as well. However, a main problem with the iterative PCA of Kiers (1997) is overfitting, i.e., while IPCA gives a small fitting error, its prediction is poor. This problem would become more serious when the amount of

✉ Vahid Nassiri
vahid.nassiri@kuleuven.be

[1] KU Leuven, BioStat, 3000 Leuven, Belgium

[2] Universiteit Hasselt, Biostat, 3590 Diepenbeek, Belgium

missing data becomes larger. Authors like Josse et al. (2009) and Ilin and Raiko (2010) proposed a regularized version of iterative PCA to overcome this problem. In this approach, a regularization parameter can control the overfitting, smaller values of this parameter would produce results similar to IPCA and larger values would regularize more. However, performance of this method depends on properly tuning this regularization parameter.

Recently, authors like Josse, Husson, and Pagés (2011), Dray and Josse (2015), Lorenzo-Seva and Van Ginkel (2016), and McNeish (2016) have considered multiple imputation (MI) in the sense of Rubin (Rubin, 2004; Schafer, 1997; Carpenter & Kenward, 2012) to deal with the missing data problem in PCA and EFA. Rubin's multiple imputation first imputes the data using, for example, a joint (Schafer, 1997) or conditional (Van Buuren, 2007) model, then in the second step performs the usual analysis on each completed (imputed) data set. The third and last step uses appropriate combination rules to combine the results from each imputed data set. An appropriate combination rule needs to respect the fact that the imputed data are, after all, unobserved. Thus, it needs to take into account that each missing value is replaced with several plausible values. The focus of the current paper is on this last approach, where the multiple imputation is done prior to the desired analysis, e.g., PCA or EFA.

In this paper, the above problem will be described and difficulties of using MI in case of factor analysis and PCA will be discussed. Possible solutions will be reviewed and an alternative simple solution will be presented. An extensive simulation study will evaluate the proposed method. Also, considering the "Divorce in Flanders" (Mortelmans et al., 2012) dataset as a case study, the application of the proposed method will be illustrated. The paper ends with concluding notes.

Using multiple imputation and exploratory factor analysis: a review

Consider p correlated random variables $\mathbf{X}_\mathbf{i}' = \{X_{i1}, \ldots, X_{ip}\}$ with observed covariance matrix $\mathrm{Cov}(\mathbf{X})$, of which the population value is Σ. In general, the idea of principal component analysis is to find as few as possible (uncorrelated) linear combinations of elements of $\mathbf{X}$ such that their variance becomes as large as possible (needed) subject to proper standardization. One can show (Johnson & Wichern, 1992) if $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ are the ordered eigenvalues of Σ, and (λ_i, e_i) is the ith eigenvalue–eigenvector pair, then the ith principal component is given by:

$$Y_i = {e_i}'\mathbf{X},\ \mathrm{Var}(Y_i) = \lambda_i,\ \mathrm{Cov}(Y_i, Y_k) = 0,\ i \neq k. \qquad (1)$$

As one may see in Eq. 1, the PCA (or exploratory factor analysis from a wider perspective) are obtained using singular value decomposition of the covariance (correlation) matrix. The eigenvalues of Σ are the roots of the characteristic polynomial $|\Sigma - \lambda I|$ where $|A|$ denotes the determinant of matrix A and I is the identity matrix of the same order as Σ. Obviously, there is no natural ordering among roots of a polynomial. Further, as one may see in Eq. 1, λ_i represents a variance. It thus makes sense to order the eigenvalues in a descending manner. Problems arise when using multiple imputation prior to the exploratory factor analysis or PCA.

Consider $X_{ij}^{(m)}$ as the observed value for ith subject $(i = 1, \ldots, N)$ of jth variable $(j = 1, \ldots, p)$ in the mth imputed dataset. The eigenvector corresponding to the largest eigenvalue of $\Sigma^{(m)} = \mathrm{Cov}(X^{(m)})$ gives the structure related to the first latent factor, and there is no guarantee that the eigenvector corresponding to the largest eigenvalue of $\Sigma^{(k)} = \mathrm{Cov}(X^{(k)})$ $(k \neq m)$ is comparable with the one from the mth imputed dataset. In other words, averaging the eigenvectors (principal axes, factor loadings) using the order or the obtained eigenvalues of the covariance matrix estimated from each imputed set is likely to lead to misleading or meaningless results.

Another difficulty of using MI prior to EFA is determining number of factors. While it is necessary to determine a common number of factors across imputed sets of data, there is no guarantee that different methods of determining number of factors would propose the same decision for each and every sets of imputed data.

In order to overcome these problems, Dray and Josse (2015) have averaged the imputed values to have one single complete dataset. Other authors like Josse et al. (2011) and Lorenzo-Seva and Van Ginkel (2016) proposed to first impute the data, then perform the PCA or factor analysis on each imputed dataset separately. After obtaining the eigenvectors (factor loadings), because of the discussed problem of ordering, one needs an intermediate step before the usual averaging. This step consists of the use of the class of generalized procrustes rotations (Ten Berge, 1977) to make these matrices as similar as possible. After rotating the obtained factor loadings simultaneously, the next step would be the usual averaging. McNeish (2016) has simply generated one set of imputed data to prevent the consequences we have discussed.

One intermediate solution in place of averaging the imputed values (Dray & Josse, 2015) or averaging the factor loadings (Lorenzo-Seva and Van Ginkel, 2016) could be to estimate the covariance matrix from imputed sets of data using Rubin's rules first, and then apply the PCA or exploratory factor analysis on this combined covariance matrix. This proposal will be discussed in the next section.

Using multiple imputation with factor analysis: a proposal

Consider $X^{(\dot{obs})}$ a dataset with missing values and $X^{(1)}, \ldots, X^{(M)}$ as the M imputed datasets with estimated covariance matrices $\widehat{\Sigma^{(1)}}, \ldots, \widehat{\Sigma^{(M)}}$. Using Rubin's rules (Rubin, 2004) the multiple imputation estimate of Σ can be obtained as follows:

$$\widetilde{\Sigma} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\Sigma_M}. \tag{2}$$

Having $\widetilde{\Sigma}$, one may perform PCA or EFA directly on it, and then the problem of factor ordering as well as determining the number of factors across imputations would vanish. Of course, that would not come for free. Estimating the covariance matrix first, and then performing EFA would make impossible the direct use of Rubin's combination rules for precision purposes. Therefore, one may consider indirect or alternative solutions. We will consider this point in more detail. It is worth noting that this is not required if no precision estimates are needed, e.g., when principal components are merely calculated for descriptive purposes or, in general, when there is no need to either make inferences about them or select a sufficiently small subset of principal components.

As was mentioned earlier, an important aspect of PCA or EFA is determining the number of factors or principal axes. While this is an important step in EFA or PCA, it is very well possible that different imputed sets of data suggest different decisions. This problem is also reported in McNeish (2016). While our proposed approach does not suffer from this problem, but still it is an important problem to address when considering MI and EFA.

One popular criterion to determine the number of factors/PCs is the proportion of explained variance. The proportion of explained variance based on the first k factors, γ_k, is:

$$\gamma_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}, \tag{3}$$

with λ_j as in Eq. 1. When using MI, one needs to ensure the correct amount of information present in the data is used. This is also important when estimating γ_k. This can be done by constructing a confidence interval for γ_k using the estimate and variance obtained by properly taking imputation process into account.

In general, even for exploratory factor analysis, it could be more informative to use an interval estimate rather than only a single point estimate to make a decision about, for example, the number of factors. Especially, when the nice properties of such point estimates (unbiasedness, consistency, etc.) are asymptotic. As Larsen and Warne (2010) also mentioned, another reason that encourages us to use confidence intervals in exploratory factor analysis is the American Psychological Association's emphasis on reporting CI's in their publication manual (APA, 2010). Larsen and Warne (2010) proposed CI's for the eigenvalues, here we extend their work and derive confidence intervals for the proportion of explained variance. This can be used to determine the common number of factors across imputed sets of data, while properly taking the imputation process into account. However, the idea is general and can be used for complete data, or other methods for analyzing incomplete data.

Consider $\Lambda = (\lambda_1, \ldots, \lambda_p)$ and Δ a diagonal matrix with $\lambda_1 \geq \ldots \geq \lambda_p$ as its diagonal elements. For large samples, we have (Anderson, 1963; Johnson & Wichern, 1992; Larsen & Warne, 2010):

$$\widehat{\Lambda} \sim N_p \left(\Lambda, \frac{2}{N} \Delta^2 \right). \tag{4}$$

Consider $\left(\widehat{\Lambda}_i, \mathrm{cov}(\widehat{\Lambda}_i)\right)$, the estimated eigenvalues and its variance from the ith imputed dataset ($i = 1, \ldots, M$). Using Rubin's rules (Rubin, 2004), the combined estimates of eigenvalues and their covariance matrix $\left(\widetilde{\Lambda}, \mathrm{cov}(\widetilde{\Lambda})\right)$ can be estimated as follows:

$$\begin{cases} \widetilde{\Lambda} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\Lambda}_i, \\ \mathrm{cov}(\widetilde{\Lambda}) = \frac{1}{M} \sum_{i=1}^{M} \mathrm{cov}(\widehat{\Lambda}_i) + \left(\frac{M+1}{M}\right) \widehat{B}, \\ \widehat{B} = \frac{1}{M-1} \sum_{i=1}^{M} (\widehat{\Lambda}_i - \widetilde{\Lambda})(\widehat{\Lambda}_i - \widetilde{\Lambda})'. \end{cases} \tag{5}$$

For the proportion of explained variance, consider the pair $(\sum_{j=1}^{k} \widetilde{\lambda_j}, \sum_{j=1}^{p} \widetilde{\lambda_j})$. Using (4) and (5) leads to:

$$\mathrm{cov} \begin{pmatrix} \sum_{j=1}^{k} \widetilde{\lambda_j} \\ \sum_{j=1}^{p} \widetilde{\lambda_j} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{pmatrix}, \tag{6}$$

with:

$$\begin{cases} \sigma_{11} = A_1 \mathrm{cov}(\widetilde{\Lambda}) A_1', \\ \sigma_{22} = A \mathrm{cov}(\widetilde{\Lambda}) A', \\ \sigma_{12} = A_1 \mathrm{cov}(\widetilde{\Lambda}) A', \end{cases} \tag{7}$$

where A is an all-one row vector of size p and A_1 as a row vector of size p with its first k elements equal to 1 and the rest $p - k$ elements equal to zero.

Now equipped with estimate of $(\sum_{j=1}^{k} \widetilde{\lambda_j}, \sum_{j=1}^{p} \widetilde{\lambda_j})$ and its covariance matrix, one needs to construct a confidence interval for their ratio. That can be used to determine a common number of factors across imputations. For constructing a confidence interval of ratio of two possible

correlated random variables, one can use Fieller's theorem (Fieller, 1954). Fieller's confidence interval for (3) can be calculated as follows:

$$C_1^2 = \frac{\sigma_{11}}{\left(\sum_{j=1}^{k} \widetilde{\lambda}_j\right)^2},\ C_2^2 = \frac{\sigma_{11}}{\left(\sum_{j=1}^{p} \widetilde{\lambda}_j\right)^2},\ r = \frac{\sigma_{11}}{\sqrt{\sigma_{11}\sigma_{22}}},$$

$$A = C_1^2 + C_2^2 - 2rC_1C_2,\ B = z_{\alpha/2}^2 C_1^2 C_2^2 (1 - r^2),$$

$$L = \widetilde{\gamma}_k \frac{1 - z_{\alpha/2}^2 rC_1C_2 - z_{\alpha/2}\sqrt{A - B}}{1 - z_{\alpha/2}^2 C_2^2}, \tag{8}$$

$$U = \widetilde{\gamma}_k \frac{1 - z_{\alpha/2}^2 rC_1C_2 + z_{\alpha/2}\sqrt{A - B}}{1 - z_{\alpha/2}^2 C_2^2}. \tag{9}$$

Note that the confidence limits in Eqs. 8 and 9 are general and one can replace $\widetilde{\Lambda}$ and $\text{cov}(\widetilde{\Lambda})$ with an estimate and its covariance obtained from any other method, and all of the equations are still valid.

The results in Eq. 4 are only valid for large samples. For small samples or where the normality assumption is violated, we propose a bootstrap confidence interval (Efron & Tibshirani, 1994). Based on Shao and Sitter (1996), we propose the following procedure to construct a bootstrap confidence interval for the proportion of explained variance for incomplete data using multiple imputation:

1. Take a bootstrap sample (a sample with replacement of size N, the same size as in the original data) from the incomplete data.
2. Impute this incomplete sample only ONCE using a predictive model.
3. Estimate the covariance matrix of the imputed data, perform EFA and compute the proportion of explained variance.
4. Repeat 1–3, e.g., 1000 times.
5. The $100(1-\alpha)\%$ confidence interval follows from the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrapped proportions of explained variance."

While the main use of constructing confidence intervals for the proportion of explained variance is to select the number of factors, it can be used for other purposes as well. It is well known (Harville, 1997) that $\text{tr}(\sum_{m=1}^{M} \Sigma_m) = \sum_{m=1}^{M} \text{tr}(\Sigma_m)$ where tr denotes the trace of the matrix. As $\text{tr}(A_n) = \sum_{j=1}^{p} \lambda_j$, where λ_j $(j = 1, \ldots, p)$ are the eigenvalues, by estimating $\widetilde{\Sigma}$ using (2), the sum of all of the eigenvalues would not change. Unfortunately, however, as Fan (1949) has shown, for matrices A, B, and $C = A + B$, with eigenvalues α_i, β_i, δ_i in descending order, respectively, that for any k, $1 \le k \le p$, we have:

$$\sum_{i=1}^{k} \delta_i \le \sum_{i=1}^{k} \alpha_i + \sum_{i=1}^{k} \beta_i. \tag{10}$$

This would mean that the proportion of explained variance obtained using eigenvalues of $\widetilde{\Sigma}$ in Eq. 2 is always smaller than $\widetilde{\gamma}_k = \sum_{j=1}^{k} \widetilde{\lambda}_j / \sum_{j=1}^{p} \widetilde{\lambda}_j$. In case that the proportion of explained variance obtained from $\widetilde{\Sigma}$ is far out of the computed CI, one needs to use our proposed method cautiously.

This phenomenon points out that either using $\widetilde{\Sigma}$, one may explain a much smaller proportion of variance with the same number of factors compared with averaging the factor loadings directly. Or this would suggest, for the given k, that the set of selected factors across the imputations are different, i.e., no matter the order, different sets of factors are selected in some of the imputed datasets. At any rate, in case of such occurrence we suggest to try other approaches as a sensitivity analysis.

Note that, the CI computed here is for $\sum_{i=1}^{M} \widehat{\gamma}_k$ and not for $\widetilde{\gamma}_k$. Therefore, using such CIs to determine the number of factors with our proposed approach is beneficial to study the validity, etc., but it is not necessary. However, this will be necessary when users want to pool the factor loadings directly, because in that case determining the common number of factors is an important issue.

The proposed method in this section together with Fieller and bootstrap confidence intervals are implemented in R package `mifa` (mifa, 2017). The documentation prepared for this package gives explanations and examples on using it. One can use this package to impute the incomplete data using fully conditional specification approach (Van Buuren, 2007), then estimate the covariance matrix using the imputed data. Also, to construct Fieller and bootstrap CIs for the proportion of explained variance for given numbers of factors. The information on how to install it is presented in the Appendix.

Simulations

In order to evaluate the proposed methodology, an extensive simulation is performed in this section. The simulation consists of three main steps. 1- generating the data, 2- creating missing values, 3- analyzing the data. These three steps were replicated 1000 times. Let's briefly go over each step.

Table 1 Summary (mean, median, and standard deviation (SD) over 1000 replications) of proportion of observed part of generated samples for different scenarios over 1000 replications

Missing	Mechanism	Mean	Median	SD
Small	MCAR	0.950	0.950	0.006
	MNAR	0.947	0.950	0.030
Large	MCAR	0.700	0.700	0.012
	MNAR	0.693	0.696	0.057

Table 2 Proportion of times (out of 1000 replications) each method led to a positive definite covariance estimate

	Small		Large	
	MCAR	MNAR	MCAR	MNAR
Listwise	1.000	0.978	0.000	0.833
Pairwise	0.518	0.926	0.002	0.153
EM	1.000	1.000	1.000	1.000
FIML	0.518	0.926	0.002	0.153
Regularized IPCA	1.000	1.000	1.000	1.000
MI	1.000	1.000	1.000	1.000

In order to generate the data, first of all a covariance matrix was generated by solving the inverse eigenvalue problem, i.e., for a given set of eigenvalues, a covariance matrix was generated using `eigeninv` package in R. The eigenvalues vector used for this simulation is as follows:

$$\Lambda = (50.0, 48.0, 45.0, 25.0, 20.0, 10.0, 5.0, 5.0, 1.0, 1.0, 0.5, 0.5, 0.5, 0.1, 0.1). \quad (11)$$

After generating the covariance matrix, its Cholesky decomposition was used to produce a set of correlated items. The sample size is set to $N = 100$ and the number of items were $p = 15$.

For each simulated dataset, two missing data mechanisms were applied from the extreme categories: non-monotone missing completely at random (MCAR) and monotone missing not at random (MNAR). Consider X_{ij} the ith observation for the jth item. For creating a non-monotone MCAR mechanism, if a random number generated from Uniform(0,1) becomes smaller than a predefined $p_{miss,MCAR}$, then X_{ij} is set as missing. For a small amount of missing data $p_{miss,MCAR}$ is set to 0.05, for a large amount of missing data it is set to 0.3. For monotone MNAR the $p_{miss,MNAR}$ is computed as $\exp(\alpha + \beta X_{ij})/\left(1 + \exp(\alpha + \beta X_{ij})\right)$ for $\beta = 1$, and $\alpha = -10$ in case of a large amount of missing data, and $\alpha = -13$ for generating a small amount of missing data. If a random number generated from a Uniform(0,1) becomes smaller than $p_{miss,MNAR}$, then X_{lj}'s for $l = i, \ldots, N$ are set as missing. The averaged Pearson correlation (over 15 items

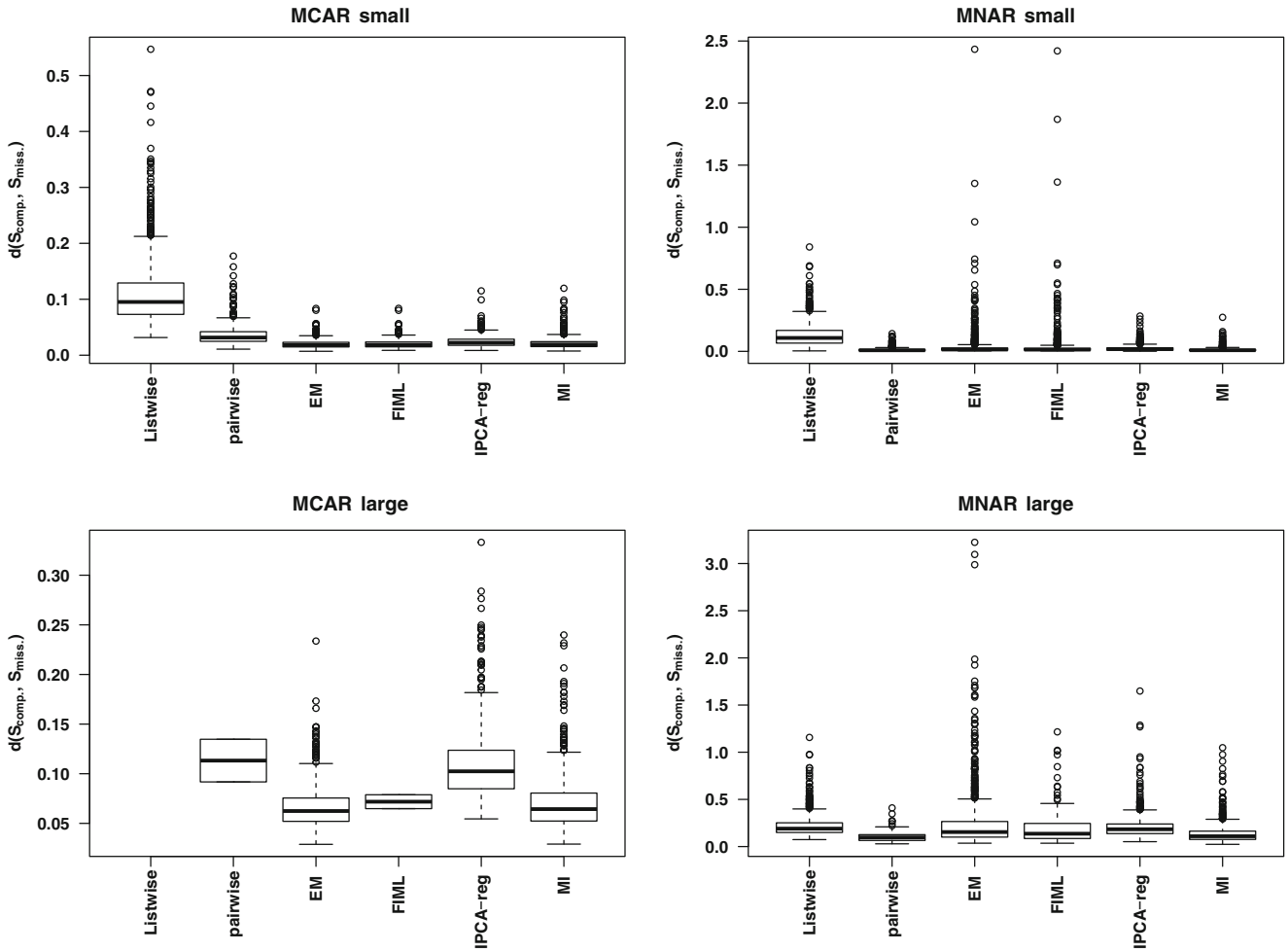

Fig. 1 Boxplots of computed distance in Eq. 12 between the estimated covariances using complete data and different methods

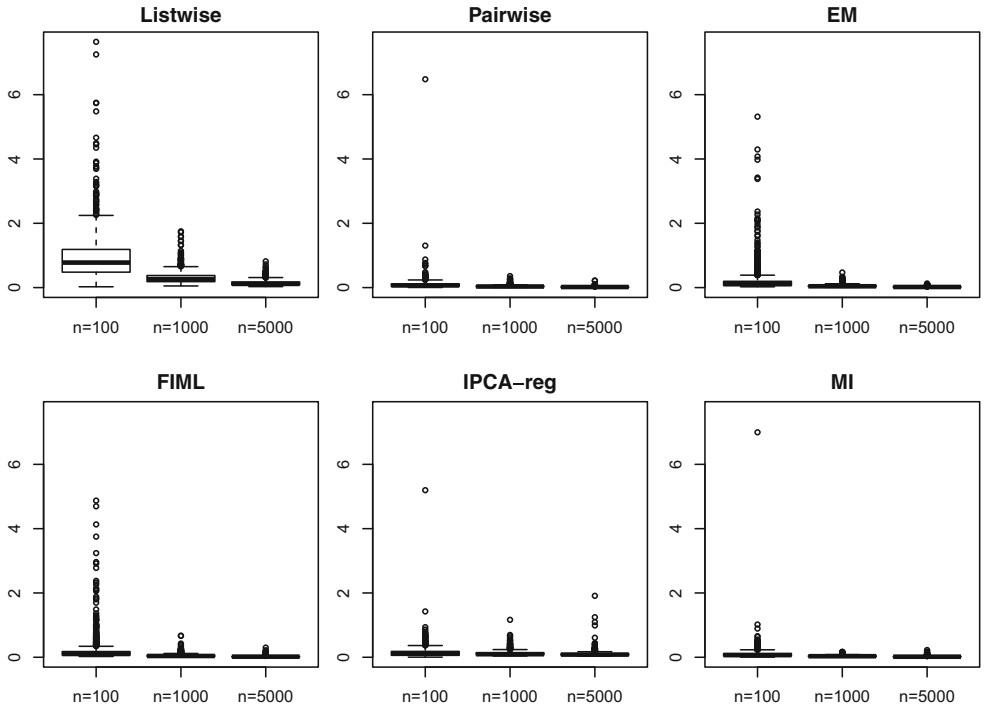

Fig. 2 Boxplots of computed distance in Eq. 12 between the estimated covariances using complete data and different methods for small of amount of MNAR missing data for $n = 100, 1000, 5000$

and 1000 replications) between the data prior to deletion and the missing data indicator is computed as 0.041 with $(-0.155, 0.234)$ 95% confidence interval for the small amount of missing data, and 0.033 $(-0.162, 0.226)$ for the large amount of missing data. Note, a correlation between a metric and a categorical variable is typically somewhat lower; our intuition stems from Pearson correlations among continuous variables. Table 1 shows a summary of proportion of observed part of sample in 1000 replications of the simulations.

The covariance matrix of incomplete data was estimated using the following methods:

1. Complete: for the complete data, before creating the missing values. This can be used as the reference to evaluate other methods.
2. Listwise deletion: this method ignores every row with missing values and estimates the covariance matrix using other rows. This can be done by setting `use='complete.obs'` in the function `cov` in R base functions.
3. Pairwise: this method uses all completely observed pairs to estimate the covariance matrix. This can be done by setting `use='pairwise.complete.obs'` in R base functions.
4. EM: this method uses the EM algorithm to compute the covariance matrix. This can be done using the function `em.norm` in the R package `norm`.
5. FIML: this method uses full information maximum likelihood to compute the covariance matrix of the incomplete data. This can be done using the function `corFiml` in the R package `psych`.
6. IPCA-reg: this method uses regularized iterative principal component analysis to impute and estimate the covariance matrix of incomplete data. This can be done using `imputePCA` function in R package `missMDA`.
7. MI: this method uses our proposal, the imputation model is based on fully conditional specification (Van Buuren, 2007) implemented in R package `mice` with a function with the same name.

Note that the maximum number of iterations for both the EM algorithm and regularized IPCA was set as 1000. Also, the regularization parameter in IPCA-reg is set as 1 (the default). For MI, when the amount of missing data was small, ten imputations were considered, and for a large amount of missing data, the incomplete data was imputed 30 times. The imputation was done using the predictive

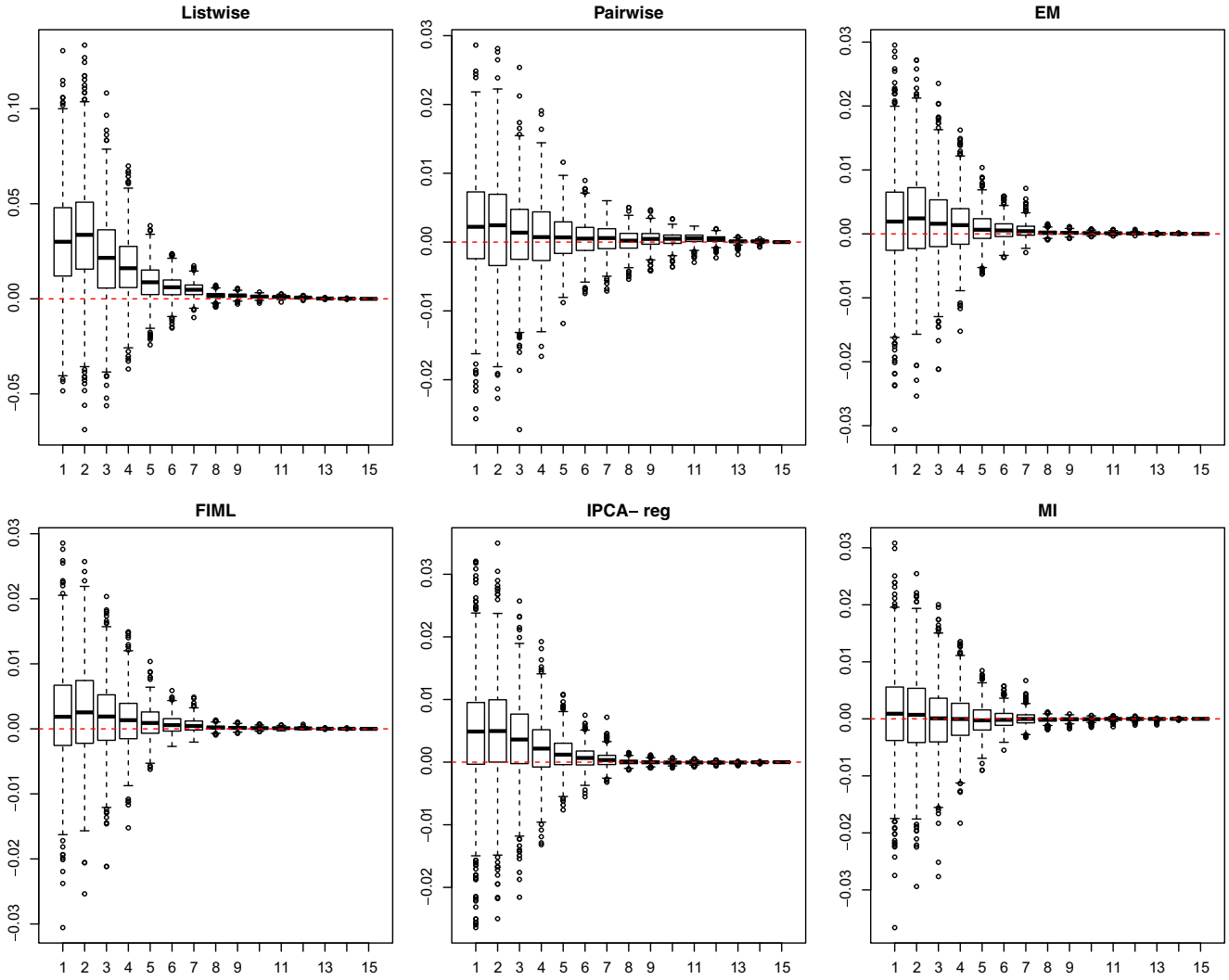

Fig. 3 Boxplots of difference of the estimated proportion of explained variance using different methods with complete data ($\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$) for MCAR- low scenario

mean matching (PMM) method (Little, 1988; Vink et al., 2014). Also, the iterations per imputation was set as 5. Furthermore, 500 sub-samples were used to construct the bootstrap CIs.

In order to summarize the results of the imputation, we have considered three main aspects: 1- the number of times each of the methods could actually lead to a positive-definite covariance estimate, 2- for cases which a positive-definite covariance was estimated, how it could be evaluated compared with the covariance estimated using complete data, 3- the proportion of explained variance in comparison with the one obtained from complete data.

Table 2 shows the proportion of times each method led to a PD covariance matrix. Note that in case of e.g., singularity, methods like listwise deletion, pairwise deletion, and FIML would lead to a non-PD covariance matrix.

For comparing the estimates obtained using each method with the one from the complete data, we use a Mahalanobis distance (Mahalanobis, 1936)-based measure $d(S_{comp.}, S_{miss.})$ as follows:

$$d(S_{comp.}, S_{miss.}) = \sqrt{\delta^T \left\{\text{Var}[\text{vech}(S_{comp.})]\right\}^{-1} \delta}, \qquad (12)$$

where $\delta = \text{vech}(S_{miss.} - S_{comp.})$ and $\text{Var}[\text{vech}(S_{comp.})] = 2(N-1)H\,S_{comp.} \otimes S_{comp.}\,H$, with H the elimination matrix, $\otimes$ the Kronecker product and vech the half vectorized version of the covariance matrix, i.e., the diagonal and

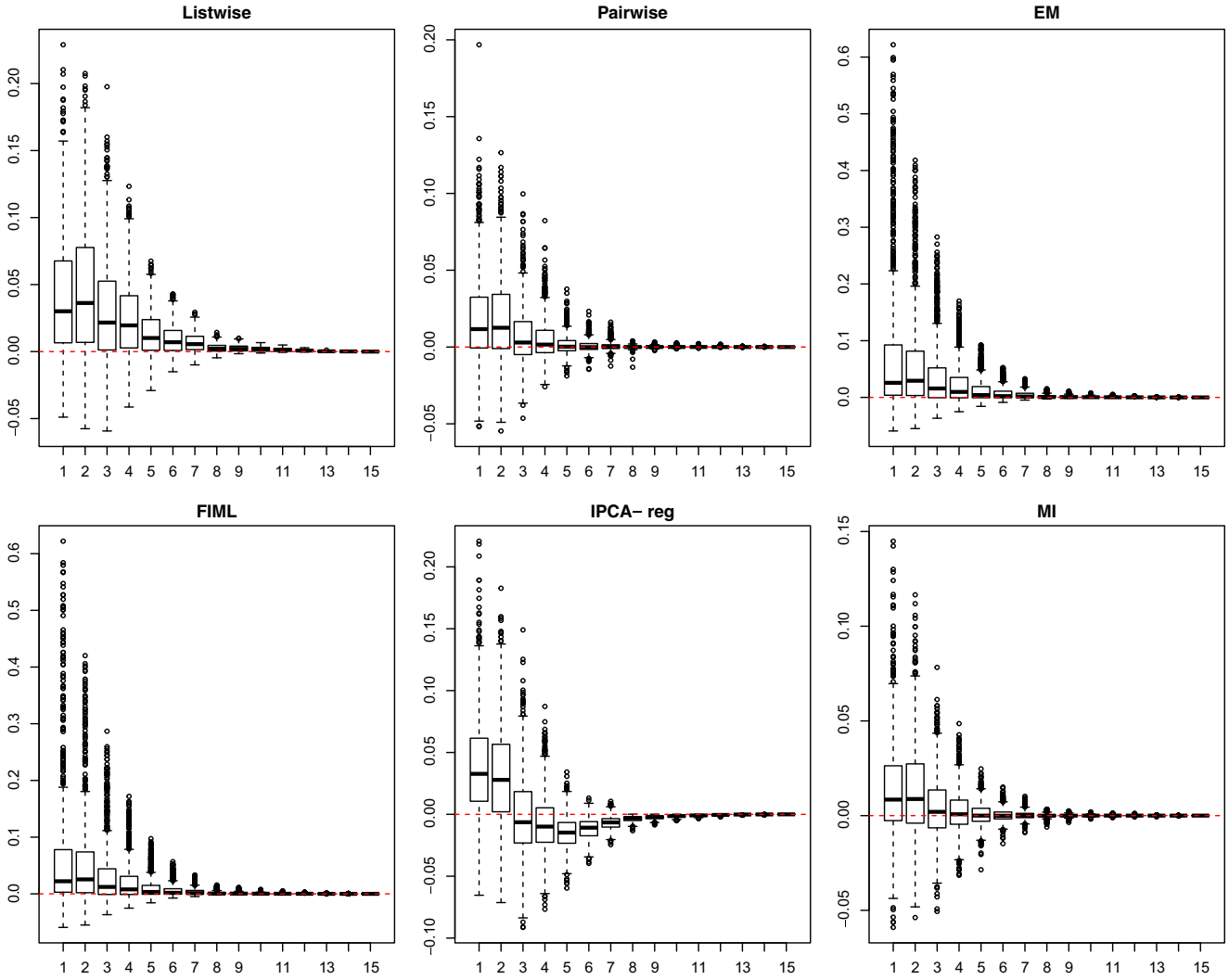

Fig. 4 Boxplots of difference of the estimated proportion of explained variance using different methods with complete data ($\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$) for MNAR- low scenario

upper (or lower) triangular elements. Figure 1 shows this measure for covariance matrices estimated using different methods compared with complete data. To see the effect of sample size, for small amounts of missing values with MNAR mechanism the simulation is repeated for $n = 100, 1000$, and 5000. Figure 2 shows (12) for covariance matrices estimated using different methods, compared with complete data for replications that such a covariance could be estimated. As one may see, for large samples, most of the methods are behaving similarly, but for small samples, which are frequent in practice, selecting an appropriate method is important.

Furthermore, for comparing the proportion of explained variance using complete data and other methods, their difference is taken as the measure: $\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$. Boxplots in Figs. 3, 4, 5 and 6 show the difference between $\widehat{\gamma}_k$ for covariance estimated using incomplete data with the one estimated using complete data with different methods in different scenarios. This is computed for all possible values of k ($k = 1, \ldots, 15$).

In addition to these, the Fieller and bootstrap confidence intervals are computed for the MI method. Tables 3 and 4 shows the averaged (over 1000 replications) of the estimated $\widehat{\gamma}_k$ and its confidence interval for $k = 1, \ldots, 10$.

As one may see in Table 2, methods like listwise deletion, pairwise deletion, and FIML would fail to estimate a positive definite covariance matrix, and the rate of this failure increases with amount of missing values. However,

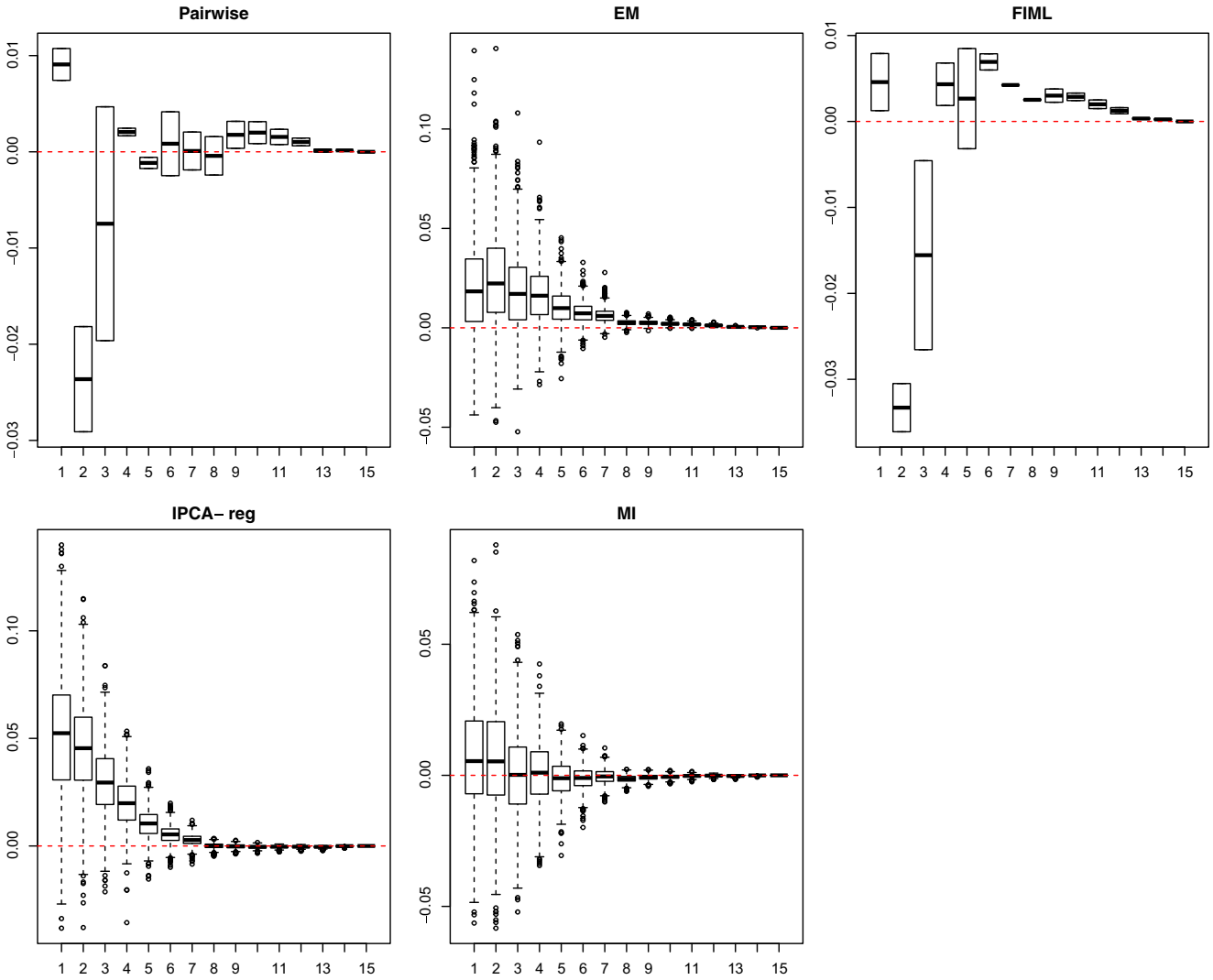

Fig. 5 Boxplots of difference of the estimated proportion of explained variance using different methods with complete data ($\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$) for MCAR- high scenario. The methods which are not presented could not be computed in none of 1000 replications

the EM algorithm, regularized IPCA, and MI would always find an estimate, though for larger amount of missing value this would take more time, i.e., more iterations or a larger number of imputations.

Figure 1 shows when the missing data are MCAR and their amount is small then almost all of the methods provide acceptable results, though, listwise deletion and pairwise deletion are not as good. In general, the EM algorithm, regularized IPCA and MI provide comparable results, though, as it is observable in Fig. 1, MI is always at least as good as its competitors.

Looking at Figs. 3, 4, 5 and 6, when it comes to estimating the proportion of explained variance, again for MCAR-small scenario, all methods perform acceptably good, especially when we look at $k = 5$ and $k = 6$, which are the desirable number of factors. However, when the mechanism becomes MNAR, or the amount of missing data increases, some biased results can be observed in comparison with estimated γ_k from the complete data. In that sense, MI has an overall better performance. Note that one would be able to get better results from regularized IPCA by tuning the regularization parameter. Here we have used the default value of the function `imputePCA` in R package `missMDA`. That could be another reason to prefer MI over regularized IPCA in this case, since no tuning is needed for MI.

In order to compare the results when using our proposal with the case where MI is directly applied on the factor

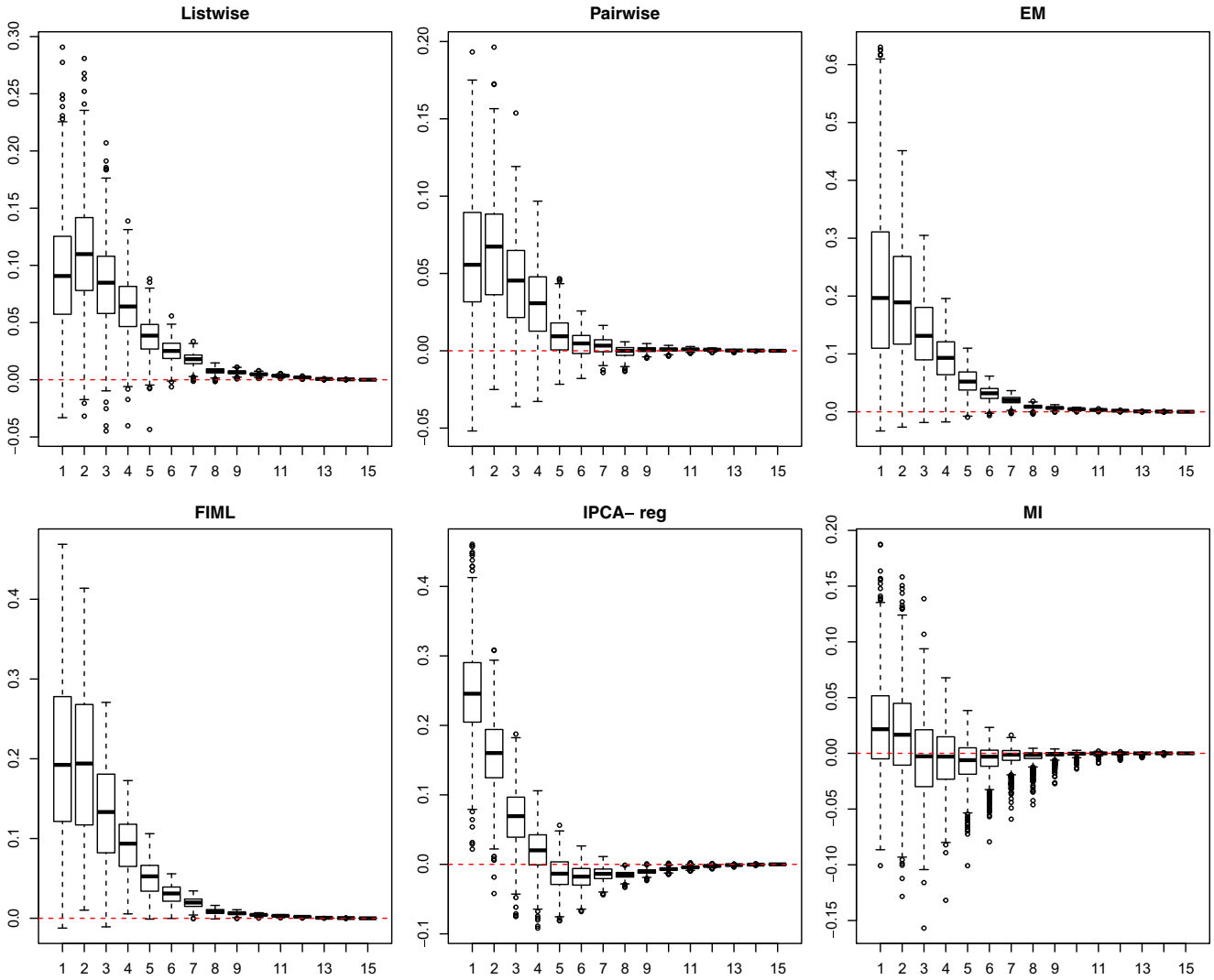

Fig. 6 Boxplots of difference of the estimated proportion of explained variance using different methods with complete data ($\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$) for MNAR- high scenario

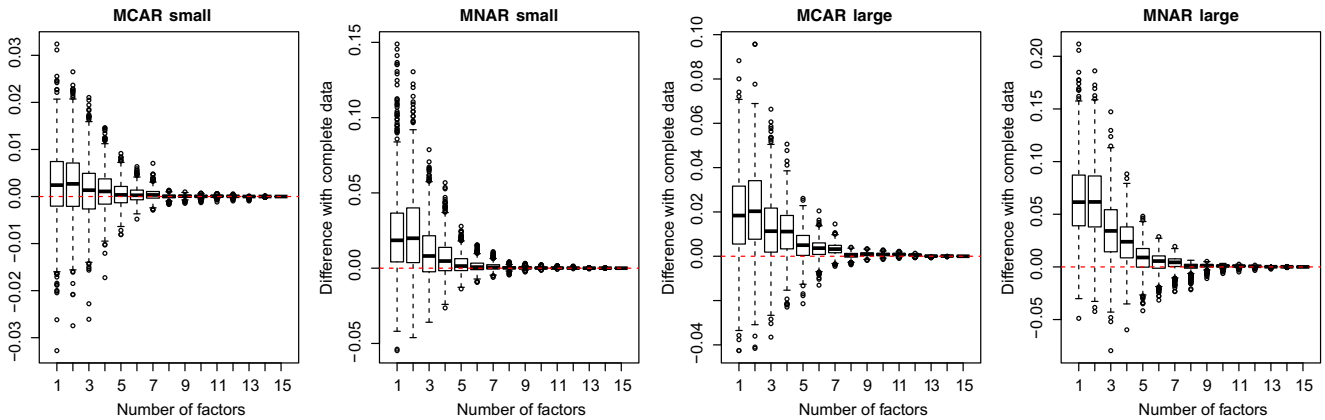

Fig. 7 Boxplots of difference of the estimated proportion of explained variance using MI directly with complete data ($\widehat{\gamma}_{k_{miss.}} - \widehat{\gamma}_{k_{comp.}}$) for different scenarios

Table 3 Small amount of missing data: the proportion of explained variance using averaged covariance, its confidence interval using Fieller's method and bootstrap for $k = 1, \ldots, 10$ and the coverage (the proportion of times each CI includes the true proportion of explained variance)

			Fieller			Bootstrap		
	k	$\widehat{\gamma_k}$	Lower	Upper	Coverage	2.5%	97.5%	Coverage
MCAR	1	0.279	0.215	0.342	1.000	0.261	0.363	0.931
	2	0.507	0.444	0.568	1.000	0.487	0.590	0.974
	3	0.691	0.642	0.737	1.000	0.667	0.752	0.998
	4	0.809	0.775	0.841	1.000	0.794	0.853	0.998
	5	0.895	0.875	0.913	1.000	0.884	0.920	1.000
	6	0.940	0.928	0.951	1.000	0.934	0.955	1.000
	7	0.965	0.957	0.972	1.000	0.962	0.975	0.993
	8	0.984	0.981	0.987	1.000	0.983	0.988	1.000
	9	0.989	0.987	0.991	1.000	0.988	0.992	0.963
	10	0.993	0.991	0.994	1.000	0.993	0.995	0.899
MNAR	1	0.292	0.223	0.375	1.000	0.272	0.413	0.861
	2	0.518	0.456	0.595	1.000	0.501	0.630	0.855
	3	0.695	0.647	0.750	1.000	0.676	0.776	0.931
	4	0.811	0.778	0.849	1.000	0.799	0.867	0.926
	5	0.895	0.876	0.916	1.000	0.886	0.927	0.986
	6	0.940	0.928	0.952	1.000	0.934	0.959	0.988
	7	0.965	0.958	0.973	1.000	0.963	0.977	0.928
	8	0.984	0.981	0.987	1.000	0.983	0.989	0.984
	9	0.989	0.987	0.991	1.000	0.989	0.993	0.886
	10	0.993	0.992	0.994	1.000	0.993	0.996	0.830

loadings, Fig. 7 shows the same boxplots for the proportion of explained variance directly obtained from imputed data. As one may see, when the amount of missingness is small, the results from both approaches are comparable, while for large amount of missing data the results of directly applying MI are generally closer to the ones obtained from complete data. Note that obtaining the corresponding eigenvectors (factor loadings) depends on using the right order among them while our approach does not suffer from this requirement.

Looking at Tables 3 and 4, we may see except for $k = 10$ in MCAR-large scenario where the estimated γ_k is slightly out of the bootstrap CI, this quantity is always in the estimated confidence interval. That would suggest: the selected set of factors are comparable across imputations, and our proposed method provides valid results.

It is also useful to see out of 1000 replications, how many times each of Fieller and bootstrap CI's contained the estimated γ_k from the complete data and the one obtained from $\widetilde{\Sigma}$. Tables 5, 6, 7 and 8 show this information for our four different scenarios. As one may see, for a small amount of missing data, for both MCAR and MNAR scenarios, the coverage of Fieller's CI for both γ_k's obtained from complete data and $\widetilde{\Sigma}$ is more than 95%. This would become smaller when the amount of missing data is large, though when k is near 5 or 6, we have almost complete coverage for MCAR and at least 80% coverage for MNAR. For all of the four scenarios, Fieller CIs are performing better than bootstrap. So that shows even with $N = 100$, the Fieller's CI performs well. One may use larger number of sub-samples to obtain better bootstrap results. Note that, although the Fieller's CIs are performing better, the bootstrap CIs are also performing acceptably fine (sometimes even better than Fieller's method, see, e.g., Table 8) when constructed for a reasonable number of factors (see Tables 5–8). Also, when the normality assumption does not hold, Fieller's method would face some difficulties. In such cases, having alternatives like bootstrap could be useful.

Divorce in Flanders

In order to illustrate the proposed methodology, we use the Divorce in Flanders (DiF) dataset (Mortelmans et al., 2012). DiF contains a sample of marriages registered in the Flemish

Table 4 Large amount of missing data: the proportion of explained variance using averaged covariance, its confidence interval using Fieller's method and bootstrap for $k = 1, \ldots, 10$ and the coverage (the proportion of times each CI includes the true proportion of explained variance)

			Fieller			Bootstrap		
	k	$\widehat{\gamma_k}$	Lower	Upper	Coverage	2.5%	97.5%	Coverage
MCAR	1	0.286	0.220	0.370	1.000	0.269	0.398	0.665
	2	0.512	0.452	0.594	1.000	0.498	0.625	0.628
	3	0.691	0.645	0.754	1.000	0.675	0.778	0.718
	4	0.810	0.780	0.855	1.000	0.801	0.873	0.630
	5	0.894	0.876	0.921	1.000	0.887	0.932	0.717
	6	0.939	0.929	0.956	1.000	0.936	0.963	0.665
	7	0.964	0.959	0.975	1.000	0.965	0.980	0.522
	8	0.983	0.980	0.988	1.000	0.982	0.990	0.747
	9	0.988	0.987	0.992	1.000	0.989	0.994	0.430
	10	0.992	0.992	0.995	0.964	0.993	0.997	0.274
MNAR	1	0.305	0.240	0.437	1.000	0.289	0.506	0.861
	2	0.524	0.478	0.648	0.990	0.520	0.703	0.855
	3	0.688	0.658	0.784	0.974	0.688	0.823	0.931
	4	0.805	0.787	0.871	0.950	0.804	0.896	0.926
	5	0.887	0.875	0.926	0.950	0.882	0.941	0.986
	6	0.934	0.928	0.958	0.945	0.930	0.967	0.988
	7	0.962	0.958	0.976	0.917	0.960	0.982	0.928
	8	0.981	0.979	0.988	0.910	0.979	0.990	0.984
	9	0.988	0.987	0.992	0.891	0.987	0.994	0.886
	10	0.992	0.992	0.995	0.842	0.992	0.997	0.830

Table 5 MCAR mechanism–small amount of missing data: the proportion of times the estimated the proportion of explained variance falls within Fieller and bootstrap confidence intervals

	Fieller				Bootstrap			
k	YY	YN	NY	NN	YY	YN	NY	NN
1	1.000	0.000	0.000	0.000	0.880	0.032	0.051	0.037
2	1.000	0.000	0.000	0.000	0.946	0.011	0.028	0.015
3	1.000	0.000	0.000	0.000	0.992	0.002	0.006	0.000
4	1.000	0.000	0.000	0.000	0.995	0.002	0.003	0.000
5	1.000	0.000	0.000	0.000	0.999	0.000	0.001	0.000
6	1.000	0.000	0.000	0.000	0.999	0.000	0.001	0.000
7	0.999	0.000	0.001	0.000	0.950	0.006	0.043	0.001
8	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
9	1.000	0.000	0.000	0.000	0.904	0.023	0.059	0.014
10	1.000	0.000	0.000	0.000	0.805	0.066	0.094	0.035
11	0.999	0.000	0.001	0.000	0.372	0.179	0.143	0.306
12	0.999	0.000	0.001	0.000	0.285	0.148	0.149	0.418
13	0.996	0.000	0.004	0.000	0.948	0.010	0.040	0.002
14	0.996	0.000	0.004	0.000	0.665	0.127	0.123	0.085
15	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000

'Y' stands for yes (the CI includes the estimated γ_k) and 'N' stands for no (the CI does not include the estimated γ_k). The *first letter* corresponds to complete data, and the *second letter* stands for our approach, e.g., YY means the estimated γ_k in both cases is included in the CI and YN means the estimated γ_k from complete data is included in the CI, but the one from $\widetilde{\Sigma}$ does not included in the CI

Table 6 MCAR mechanism–large amount of missing data: the proportion of times the estimated proportion of explained variance falls within Fieller and bootstrap confidence intervals

	Fieller				Bootstrap			
k	YY	YN	NY	NN	YY	YN	NY	NN
1	1.000	0.000	0.000	0.000	0.471	0.230	0.194	0.105
2	0.998	0.000	0.002	0.000	0.430	0.246	0.198	0.126
3	0.994	0.000	0.006	0.000	0.631	0.239	0.087	0.043
4	0.990	0.000	0.010	0.000	0.503	0.260	0.127	0.110
5	0.997	0.000	0.003	0.000	0.654	0.226	0.063	0.057
6	0.990	0.000	0.010	0.000	0.586	0.235	0.079	0.100
7	0.966	0.000	0.034	0.000	0.318	0.187	0.204	0.291
8	0.993	0.000	0.007	0.000	0.706	0.185	0.041	0.068
9	0.960	0.000	0.040	0.000	0.247	0.161	0.183	0.409
10	0.880	0.029	0.084	0.007	0.099	0.097	0.175	0.629
11	0.572	0.168	0.176	0.084	0.016	0.042	0.099	0.843
12	0.368	0.278	0.185	0.169	0.027	0.043	0.070	0.860
13	0.603	0.285	0.084	0.028	0.376	0.146	0.091	0.387
14	0.488	0.236	0.186	0.090	0.127	0.084	0.089	0.700
15	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000

'Y' stands for yes (the CI includes the estimated γ_k) and 'N' stands for no (the CI does not include the estimated γ_k). The first letter corresponds to complete data and the seconds letter stands for our approach, e.g., YY means the estimated γ_k in both cases is included in the CI and YN means the estimated γ_k from complete data is included in the CI, but the one from $\widetilde{\Sigma}$ does not included in the CI

Table 7 MNAR mechanism–small amount of missing data: the proportion of times the estimated proportion of explained variance falls within Fieller and bootstrap confidence intervals

	Fieller				Bootstrap			
k	YY	YN	NY	NN	YY	YN	NY	NN
1	0.977	0.000	0.023	0.000	0.574	0.075	0.287	0.064
2	0.971	0.000	0.029	0.000	0.530	0.075	0.325	0.070
3	0.984	0.000	0.016	0.000	0.816	0.049	0.115	0.019
4	0.978	0.000	0.022	0.000	0.787	0.050	0.139	0.024
5	0.993	0.000	0.007	0.000	0.927	0.008	0.059	0.005
6	0.993	0.000	0.007	0.000	0.930	0.008	0.058	0.003
7	0.987	0.000	0.013	0.000	0.810	0.041	0.118	0.031
8	0.984	0.000	0.016	0.000	0.914	0.014	0.070	0.002
9	0.992	0.000	0.008	0.000	0.728	0.064	0.158	0.050
10	0.984	0.000	0.016	0.000	0.663	0.084	0.167	0.086
11	0.977	0.000	0.023	0.000	0.366	0.144	0.184	0.307
12	0.965	0.000	0.035	0.000	0.283	0.144	0.163	0.410
13	0.989	0.000	0.011	0.000	0.813	0.035	0.126	0.026
14	0.980	0.000	0.019	0.001	0.642	0.073	0.158	0.127
15	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000

'Y' stands for yes (the CI includes the estimated γ_k) and 'N' stands for no (the CI does not include the estimated γ_k). The *first letter* corresponds to complete data and the *second letter* stands for our approach, e.g., YY means the estimated γ_k in both cases is included in the CI and YN means the estimated γ_k from complete data is included in the CI, but the one from $\widetilde{\Sigma}$ is not included in the CI

Table 8 MNAR mechanism–large amount of missing data: the proportion of times the estimated proportion of explained variance falls within Fieller and bootstrap confidence intervals

	Fieller				Bootstrap			
k	YY	YN	NY	NN	YY	YN	NY	NN
1	0.902	0.000	0.098	0.000	0.574	0.075	0.287	0.064
2	0.776	0.007	0.214	0.003	0.530	0.075	0.325	0.070
3	0.819	0.024	0.155	0.002	0.816	0.049	0.115	0.019
4	0.770	0.045	0.180	0.005	0.787	0.050	0.139	0.024
5	0.831	0.049	0.119	0.001	0.927	0.008	0.059	0.005
6	0.813	0.049	0.132	0.006	0.930	0.008	0.058	0.003
7	0.733	0.069	0.184	0.014	0.810	0.041	0.118	0.031
8	0.766	0.058	0.144	0.032	0.914	0.014	0.070	0.002
9	0.703	0.084	0.188	0.025	0.728	0.064	0.158	0.050
10	0.633	0.121	0.209	0.037	0.663	0.084	0.167	0.086
11	0.495	0.149	0.283	0.073	0.366	0.144	0.184	0.307
12	0.432	0.147	0.277	0.144	0.283	0.144	0.163	0.410
13	0.596	0.102	0.245	0.057	0.813	0.035	0.126	0.026
14	0.608	0.033	0.291	0.068	0.642	0.073	0.158	0.127
15	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000

'Y' stands for yes (the CI includes the estimated γ_k) and 'N' stands for no (the CI does not include the estimated γ_k). The *first letter* corresponds to complete data and the *second letter* stands for our approach, e.g., YY means the estimated γ_k in both cases is included in the CI and YN means the estimated γ_k from complete data is included in the CI, but the one from $\widetilde{\Sigma}$ does not included in the CI

Region of Belgium, between 1971 and 2008 with an oversampling for dissolved marriages (2/3 dissolved and 1/3 intact marriages). As part of this study, the participants were asked to complete the validated Dutch version (Denissen et al., 2008) of the Big Five Inventory (BFI) (John & Srivastava, 1999). The validity of the BFI for DiF data is investigated in Lovik et al. (2017).

The sample at hand consists of 9385 persons in 4472 families. In each family, mother, father, step mother, step father, and one child over 14 were asked to fill in the BFI. Note that, depending on the presented family roles and the number of people that agreed to participate, the size of the families could vary between 1 and 5. Among these 9385 persons, there are 1218 persons with at least one non-response (out of 44 items). As our main purpose here was to illustrate the use of the proposed method, in order to get rid of the problem of intra-correlation within the families, one person from each family was selected at random to form a sample of uncorrelated subjects. As a result, a random sample of size 4472 was taken where 515 of them had at least one non-response.

This incomplete dataset was imputed using the fully conditional specification (FCS) of Van Buuren (2007) using the `MICE` package in R (Buuren & Groothuis-Oudshoorn, 2011) with PMM method. The imputation was done $M = 25$ times. The covariance matrix was estimated for each of the imputed sets of data and the exploratory factor analysis was done on the averaged estimated covariance matrix, as well as on each imputed set. The latter was used to construct a confidence interval for the proportion of explained variance. This can be done using function `mifa.cov` in the R package `mifa`, available at (mifa, 2017), as `mifa.cov(data.sub, n.factor=5, M=25, maxit.mi=10, method.mi='pmm',alpha=0.05, rep.boot=1000, ci=TRUE)`. `data.sub` here is the selected incomplete sub-sample.

The estimated factor loadings are presented in Table 9. The Fieller's confidence interval for the proportion of explained variance using five factors is obtained as $(0.428, 0.439)$. Also, the bootstrap CI is obtained as $(0.429, 0.441)$. The estimated proportion of explained variance of the first five factors using the proposed methodology is 0.434, which falls within both of the estimated intervals. This is coherent with the validity of the proposed methodology for this dataset as in Lovik et al. (2017).

Table 9 Factor loadings using oblimin rotation of DiF data using the estimated covariance matrix from multiply imputed data using $M = 25$ imputations

English items*	Factor loadings				
	O	C	N	E	A
19. worries a lot	0.040	0.052	**0.664**	−0.097	0.067
14. can be tense	0.084	0.126	**0.647**	−0.068	−0.049
9r**. is relaxed, handles stress well	−0.299	−0.063	**0.551**	−0.030	0.004
39. gets nervous easily	−0.021	−0.004	**0.701**	0.000	−0.017
24r. is emotionally stable, not easily upset	−0.285	−0.100	**0.458**	0.001	−0.005
34r. remains calm in tense situations	−0.286	−0.128	**0.518**	0.063	−0.064
4. is depressed, blue	0.018	−0.046	**0.394**	−0.260	−0.097
29. can be moody	0.167	0.025	**0.388**	0.004	−0.306
1. is talkative	0.132	0.014	0.110	**0.609**	0.049
21r. tends to be quiet	−0.087	−0.064	−0.018	**0.730**	−0.041
16. generates a lot of enthusiasm	0.381	0.139	0.015	**0.418**	0.211
36. is outgoing, sociable	0.186	0.004	0.117	**0.431**	0.415
6r. is reserved	−0.111	−0.009	−0.143	**0.630**	0.029
31r. is sometimes shy	−0.183	0.053	−0.230	**0.556**	−0.062
11. is full of energy	0.300	0.253	−0.181	**0.301**	0.029
26. has an assertive personality	0.237	0.248	−0.146	**0.310**	−0.143
40. likes to reflect, play with ideas	**0.540**	0.231	−0.005	−0.041	−0.009
25. is inventive	**0.582**	0.179	−0.149	0.065	−0.030
30. values artistic, aesthetic experiences	**0.558**	−0.054	0.017	−0.182	0.151
5. is original, comes up with new ideas	**0.515**	0.122	−0.053	0.113	−0.032
15. is ingenious, a deep thinker	**0.425**	0.325	0.115	−0.041	−0.078
20. has an active imagination	**0.537**	−0.141	0.041	0.102	−0.000
10. is curious about many different things	**0.498**	0.137	−0.084	0.141	0.017
44. is sophisticated in art, music, or literature	**0.432**	−0.135	−0.035	−0.103	0.088
41r. has few artistic interests	**0.348**	−0.112	−0.104	−0.131	0.116
35r. prefers work that is routine	**0.138**	−0.042	−0.188	0.006	−0.087
3. does a thorough job	0.127	**0.571**	0.059	0.029	−0.042
28. perseveres until the task is finished	0.109	**0.642**	0.026	−0.035	0.029
18r. tends to be disorganized	−0.379	**0.563**	−0.004	−0.019	0.061
23r. tends to be lazy	−0.263	**0.529**	−0.025	0.037	0.136
13. is a reliable worker	0.135	**0.478**	0.070	0.049	0.089
33. does things efficiently	0.168	**0.616**	−0.026	−0.014	0.076
38. makes plans and follows through with them	0.233	**0.524**	−0.045	0.147	−0.049
43r. is easily distracted	−0.220	**0.448**	−0.290	−0.039	0.021
8r. can be somewhat careless	−0.400	**0.457**	0.022	−0.039	0.104
32. is considerate and kind to almost everyone	0.193	0.108	0.126	0.064	**0.534**
17. has a forgiving nature	0.162	0.025	0.036	0.044	**0.446**
7. is helpful and unselfish with others	0.171	0.147	0.060	−0.011	**0.217**
12r. starts quarrels with others	−0.071	0.058	−0.288	−0.132	**0.389**
37r. is sometimes rude to others	−0.187	0.091	−0.179	−0.165	**0.522**
27r. can be cold and aloof	−0.138	−0.002	−0.063	0.205	**0.479**
22. is generally trusting	0.200	−0.096	−0.015	0.043	**0.344**
2r. tends to find fault with others	−0.195	0.003	−0.249	−0.181	**0.380**
42. likes to cooperate with others	0.173	0.103	0.025	0.229	**0.303**

*The English translations are taken from Denissen et al. (2008)

**Negatively framed items were reversed before analysis

N = Neuroticism, E = Extraversion, O = Openness to Experience,

C = Conscientiousness, A = Agreeableness.

The bold emphasis indicates the "main items (primary loadings)" related to each factor.

Conclusions

Nonresponse and missing values are among at the same time major and common problems in data analysis, especially when it comes to survey data. Multiple imputation, which was first introduced to deal with nonresponse in surveys (Rubin, 2004), has become a key and effective tool for dealing with this problem. While MI has become a very commonly used approach to handle missing data in medical sciences, its use in psychology is increasing as well. As Lorenzo-Seva and Van Ginkel (2016) mentioned, a Google search for the terms *psychology "multiple imputation"* produced about 131,000 hits. Repeating it now, the number of hits has increased to 171,000. This shows the growing use of MI in the field of psychology and psychometry, hence the necessity to develop frameworks for using MI, in conjunction with various methods commonly used in psychological research. However, when it comes to combining this methodology with techniques like exploratory factor analysis and principal component analysis, due to the problems of determining a common number of factors/principal components and then ordering them, combining the results from different sets of imputed data becomes as issue.

This problem is addressed in this article and a pragmatic solution is proposed, which is justified by theoretical discussion and reasoning. Our proposal states to first estimate the covariance matrix of the correlated variables and then perform the EFA/PCA on this single matrix. The theoretical aspects of this methodology are studied and investigated. As an extension of the work of Larsen and Warne (2010), confidence intervals are proposed for the proportion of explained variance, which can be used to determine the common number of factors across imputations. Also, such confidence intervals can be useful to decide on the validity of the proposed method.

The simulation results show comparable performance of the proposed method, when compared to alternative methodologies. To evaluate our proposal in real situations, it is applied to an incomplete BFI dataset; the result was definitely acceptable. The main advantages of using the proposed methodology are: it is compatible with any imputation methodology; implementing it is very straightforward and no extra programming effort is needed. Therefore, it can be used within any desired statistical package or software. It is fast and practical. Also, the proposed confidence intervals for the proportion of explained variance can be used to determine the number of factors.

The proposed ideas in this article are also implemented in an R package `mifa`, which is available at mifa (2017). That would make it more available for practice.

Other MI-based solutions for exploratory factor analysis of incomplete data come with their own pros and cons. Dray and Josse (2015) pool imputed datasets, while this should be done for parameters. Our approach solves this issue by working at the parameter level. Also, McNeish (2016) considers only $M = 1$ imputed dataset. This does not comply with the main goal of multiple imputation, which is considering the uncertainty imposed by replacing the unobserved values by predicted values obtained based on observed values by replacing each missing value by several plausible candidates. With our proposal, it is straightforward to use $M > 1$ imputed datasets. Finally, the approach in Lorenzo-Seva and Van Ginkel (2016) does not suffer from any of the issues discussed above, but implementing it in practice is difficult. To the best of our knowledge, other than the stand-alone software FACTOR (Lorenzo-Seva & Ferrando, 2006), no other publicly available implementation of this approach exists. Ideally, the results of this article and the R implementation of our proposal would encourage the research on as well as use of MI for EFA of incomplete data.

Acknowledgements Financial support from the IAP research network # P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The research leading to these results has also received funding from the European Seventh Framework programme FP7 2007 - 2013 under grant agreement Nr. 602552. We gratefully acknowledge support from the IWT-SBO ExaScience grant. We are grateful for suggestions made by anonymous referees, which have greatly helped to improve this manuscript.

Appendix: using R package `mifa`

The package is available on I-BioStat's website (mifa, 2017). One needs to download the package's zip file. After extracting the `mifa` folder, one can install the package or simply run each function separately. The functions will be available in `..mifa/ R`. The documentation of how to use different functions can be found in `mifa.pdf`.

For installing the package, one needs to first install `Rtools` from Rtools (2017), then install the package `devtools` in R. Having these two installed, the following code can be used to install the package `mifa`:

```
devtools::install('..\\mifa')
```

Note that the input in the code above should be the path of the extracted `mifa` folder.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, *34*(1), 122–148.

APA (2010). *Publication manual of the American Psychological Association*, 6th edn. American Psychological Association Washington.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3).

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1–38.

Denissen, J. J., Geenen, R., Van Aken, M. A., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157.

Dray, S., & Josse, J. (2015). Principal component analysis with missing values: A comparative survey of methods. *Plant Ecology*, *216*(5), 657–667.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457.

Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, *35*, 652–655.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 175–185.

Gabriel, K. R., & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, *21*(4), 489–498.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.

Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, *11*, 1957–2000.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*(1999), 102–138.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* Vol. 4. Englewood Cliffs: Prentice Hall.

Josse, J., Husson, F., & Pagès, J. (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la Socié,té Française de Statistique*, *150*(2), 28–51.

Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, *5*(3), 231–246.

Kiers, H. A. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, *62*(2), 251–266.

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*, *42*(3), 871–876.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). Factor: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91.

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.

Lovik, A., Nassiri, V., Verbeke, G., Molenberghs, G., & Sodermans, A. K. (2017). Psychometric properties and comparison of different techniques for factor analysis on the Big Five Inventory from a Flemish sample. *Personality and Individual Differences*, *117*, 122–129.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49–55.

McNeish, D. (2016). Exploratory factor analysis with small samples and missing data. *Journal of Personality Assessment*, 1–16.

mifa (2017). Retrieved from https://ibiostat.be/online-resources/online-resources/expfactor

Mortelmans, D., Pasteels, I., Van Bavel, J., Bracke, P., Matthijs, K., & Van Peer, C. (2012). Divorce in Flanders. Data collection and code book. Retrieved September, 22.

Rtools (2017). Retrieved from https://cran.r-project.org/bin/windows/Rtools/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). Wiley.

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.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Shao, J., & Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*(435), 1278–1288.

Ten Berge, J. M. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, *42*(2), 267–276.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

Vink, G., Frank, L. E., Pannekoek, J., & Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90.

Wold, H., & Lyttkens, E. (1969). Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin of the International Statistical Institute*, *43*, 1.