Comparisons among several methods for handling missing data in principal component analysis (PCA)

Noname ADAC manuscript No.
(will be inserted by the editor)

Comparisons among several methods for handling missing data in principal component analysis (PCA)

Sébastien Loisel · Yoshio Takane

The date of receipt and acceptance will be inserted by the editor

Abstract Missing data are prevalent in many data analytic situations. Those in which principal component analysis (PCA) is applied are no exceptions. The performance of five methods for handling missing data in PCA is investigated, the missing data passive (MDP) method, the weighted low rank approximation (WLRA) method, the regularized PCA (RPCA) method, the trimmed scores regression (TSR) method, and the data augmentation (DA) method. Three complete data sets of varying sizes were selected, in which missing data were created randomly and non-randomly. These data were then analyzed by the five methods, and their parameter recovery capability, as measured by the mean congruence coefficient between loadings obtained from full and missing data, is compared as functions of the number of extracted components (dimensionality) and the proportion of missing data (censor rate). For randomly censored data, all five methods worked well when the dimensionality and censor rate were small. Their performance deteriorated, as the dimensionality and censor rate increased, but the speed of deterioration was distinctly faster with the WLRA method. The RPCA method worked best and the DA method came as a close second in terms of parameter recovery. However, the latter, as implemented here, was found to be extremely time-consuming. For non-randomly censored data, the recovery was also affected by the degree of non-randomness in censoring processes. Again the RPCA method worked best, maintaining good to excellent recoveries when the censor rate was small and the dimensionality of solutions was not too excessive.

Sébastien Loisel
Department of Mathematics, Heriot-Watt University, Edinburgh, EH14 4AS, UK
Phone: +44 131 451 3234 Fax: +44 131 451 3249 E-mail: sloisel@gmail.com

Yoshio Takane
Department of Psychology, University of Victoria, 5173 Del Monte Avenue, Victoria, BC V8Y 1X3, Canada
Phone: 250-744-0076 Fax: 250-744-0076 E-mail: yoshio.takane@mcgill.ca

Keywords Homogeneity criterion · Missing data passive (MDP) method · Alternating least squares (ALS) algorithm · Weighted low rank approximation (WLRA) method · Regularized PCA (RPCA) method · Trimmed scores regression (TSR) method · Data augmentation (DA) method · Congruence coefficient

Subject classification JEL C25, AMS 15A03 15A09

1 Introduction

Missing data occur frequently in many practical data analytic situations. Those in which principal component analysis (PCA) is often applied are no exceptions. Various methods have been developed to deal with missing data in PCA, ranging from simple but naive techniques such as listwise and pairwise deletions to more sophisticated but computationally more demanding techniques. The latter may further be divided into two groups, one consisting of distribution-free least squares (LS) methods, and the other based on EM algorithms (e.g., Bernaards and Sijtsma 2000; Serneels and Verdonck 2008; Stanimirova, Daszykowski, and Walczak 2007) or multiple imputations (Rubin 1987; Van Ginkel and Kroonenberg 2014) under specific distributional assumptions. See, for example, Ilin and Raiko (2010), and Van Ginkel, Kroonenberg, and Kiers (2014) for overviews of (some of) these techniques.

In this paper, we examine the performance of five representative methods of dealing with missing data in PCA under various missingness conditions. Here, the "representative" methods mean those that have been shown to work well in at least one previous simulation study (see the paragraph below), or those that have not been compared previously but have notably attractive features (e.g., non-iterative). Specifically, we compare the parameter recovery capability of the five methods as functions of the number of components extracted (dimensionality of solutions), the proportion of missing data (censor rate), and the degree of non-randomness in creating missing data. The first method we consider is called the missing data passive (MDP) method based on homogeneity analysis (Meulman 1982; Takane and Oshima-Takane 2003). The second method is the weighted low rank approximation (WLRA) method (Gabriel and Zamir 1979; Grung and Manne 1998; Walczak and Massart 2001). The third method is the regularized PCA (RPCA) method (Josse, Husson, and Pagès 2009; Josse and Husson 2012). The fourth method is the trimmed scores regression (TSR) method (Folch-Fortuny, Arteaga, and Ferrer 2015). The fifth method is a variant of multiple imputation method (Van Ginkel and Kroonenberg 2014; Van Ginkel et al. 2014) called the data augmentation (DA) method (Schafer 1997; Tanner and Wong 1987). The MDP method, the WLRA method, and the TSR method are distribution-free, while the other two (the RPCA method and the DA method) involve distributional assumptions. The MDP method is non-iterative, while the other four are iterative. In the present study, three complete real data sets of varying sizes are selected, in which missing data are created randomly and non-randomly in different

proportions. These artificially created incomplete data are then analyzed by the five methods, and their parameter recovery capability, as measured by the mean congruence coefficient (Tucker 1951), is examined against the original complete data.

Several notable simulation studies have been conducted recently (Dray and Josse 2015; Folch-Fortuny, Arteaga, and Ferrer 2015; van Ginkel et al. 2014) to compare the performance of several methods for PCA with missing values. However, none of these studies included all of the methods we consider in the present study. The TSR method (Folch-Fortuny et al. 2015) was not included in van Ginkel et al.'s study because it was proposed after van Ginkel et al.'s (2014) study. This method is interesting because it was found to work well in a wide range of situations in Folch-Fortuny et al.'s (2015) study, which in turn did not include the MDP method or the RPCA method. Van Ginkel et al. (2014) precluded the WLRA method in their study on the account that it tended to overfit missing data when too many components were extracted, and instead included its regularized version called the RPCA method. We have nonetheless chosen to include the WLRA method (as well as the RPCA method) because Folch-Fortuny et al. (2015) reported that it worked reasonably well under a variety of conditions. We have also included the DA method because it was found to work consistently well in both Folch-Fortuny et al.'s and van Ginkel et al.'s studies.

Van Ginkel et al. (2014) limited the proportion of missing data to 15%, while Dray and Josse (2015) and Folch-Fortuny et al. (2015) examined up to 50 to 90% censor rates. While the latter proportions seem too excessive in practical sense, data with more than 15% missing values can occur quite commonly. In test equating situations (Shibayama 1995), for example, data with more than 15% missing values are regularly encountered. Folch-Fortuny et al. (2015) have pointed out that while in chemometrics environments, practitioners usually deal with 5 to 20% of missing values, in complex chemical industrial processes, 30 to 60% of missing data can occur. In Big Data situations with several hundred variables, even larger proportions of missing data can arise (e.g., Ilin and Raiko, 2010). In the present study, we include missing data proportions up to 30%. We also examine the effects of weak components on the recovery.

This paper is organized as follows: In the following section (Section 2), we discuss the five methods to be compared in this paper, in which we also point out their potential advantages and disadvantages. In Section 3, we state the design of our study, introduce the performance measure (the mean congruence coefficient) to be used in comparison, describe the data sets to be used, and report the main results. In the final section, we provide a summary of the results and recommendations.

2 The five methods

We first introduce some common notations. We then discuss a method of PCA of complete data, which will later be generalized to handling missing data. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_m]$ denote an n-case by m-variable data matrix possibly with missing entries. Let $\mathbf{D}_{w_j}$ $(j = 1, \cdots, m)$ indicate a diagonal matrix whose k-th diagonal element is unity if the k-th element of $\mathbf{x}_j$ is observed and zero otherwise. Let $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_m]$ denote a columnwise standardized data matrix. The standardization is performed for each variable with resect to observed portions of the data.

Let $\mathbf{F}$ denote the n by r-component matrix of component scores, and let $\mathbf{A}$ denote the m by r matrix of component loadings. For complete data, these matrices can be derived by first obtaining the (compact) singular value decomposition (SVD) of $\mathbf{Z}$, i.e.,

$$\mathbf{Z} = \mathbf{BDC}', \tag{1}$$

where $\mathbf{B}$ is the n by t matrix of left singular vectors (where $t = \text{rank}(\mathbf{Z}) \geq r$, $\mathbf{C}$ is the m by t matrix of right singular vectors, and $\mathbf{D}$ the t by t diagonal matrix of singular values arranged in descending order. Let $\mathbf{B}_r$, $\mathbf{C}_r$, and $\mathbf{D}_r$ denote the portions of $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ corresponding to the r dominant singular values. Then, $\mathbf{F}$ and $\mathbf{A}$ are obtained by

$$\mathbf{F} = n^{1/2}\mathbf{B}_r, \tag{2}$$

and

$$\mathbf{A} = \mathbf{C}_r\mathbf{D}_r/n^{1/2}. \tag{3}$$

There are at least two distinct criteria from which the above solutions are obtained: One is to minimize

$$\phi^{(c)}(\mathbf{F}, \mathbf{U}) = \frac{1}{m}\sum_{j=1}^{m} \text{SS}(\mathbf{F} - \mathbf{z}_j\mathbf{u}_j'), \tag{4}$$

where $\mathbf{u}_j'$ is an r-element vector of weights, $\mathbf{U}' = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, and $\text{SS}(\mathbf{Y}) = \text{tr}(\mathbf{Y}'\mathbf{Y})$ for any matrix $\mathbf{Y}$. (The parenthesized superscript $^{(c)}$ on ϕ indicates that this criterion is for complete data.) This is called a homogeneity criterion, since minimizing $\phi^{(c)}$ creates $\mathbf{z}_j\mathbf{u}_j'$'s that are as homogeneous as possible over j $(j = 1, \cdots, m)$. The other criterion is

$$\tau^{(c)}(\mathbf{F}, \mathbf{A}) = \frac{1}{m}\text{SS}(\mathbf{Z} - \mathbf{FA}'), \tag{5}$$

which is called a low rank approximation criterion, since minimizing $\tau^{(c)}$ obtains the matrix of the best low rank approximation $\mathbf{FA}'$ to the standardized data matrix $\mathbf{Z}$. Both criteria are minimized subject to the normalization restriction that $\mathbf{F}'\mathbf{F} = n\mathbf{I}_r$. For complete data, these two criteria are simply related (Gifi, 1990, p. 168), that is, one criterion is a simple linear transformation of the other, and they lead to identical solutions. This, however, will not

be true when missing data exist. The first two (MDP and WLRA) of the five methods compared in this paper derive from generalizations of the two criteria above, which yield distinct solutions in the presence of missing observations.

2.1 The missing data passive (MDP) method

We derive the MDP method by generalizing (4) as follows (Gifi 1990; Takane and Oshima-Takane 2003):

$$\phi^{(ic)}(\mathbf{F},\mathbf{U}) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{SS}(\mathbf{F}-\mathbf{z}_j\mathbf{u}_j')_{D_{w_j}}, \tag{6}$$

where $\mathrm{SS}(\mathbf{Y})_W = \mathrm{tr}(\mathbf{Y}'\mathbf{W}\mathbf{Y})$ for any matrix $\mathbf{Y}$ and a symmetric *nnd* (non-negative definite) matrix $\mathbf{W}$. (The parenthesized superscript $^{(ic)}$ stands for incomplete data.) A slightly generalized version of the above criterion where the vector $\mathbf{z}_j$ is replaced by a matrix of dummy variables was first used for missing data in multiple correspondence analysis (Meulman 1982). This criterion is minimized with respect to $\mathbf{F}$ and $\mathbf{U}$ under the restriction that

$$\mathbf{F}'\mathbf{D}_w\mathbf{F} = n\mathbf{I}_r, \tag{7}$$

where

$$\mathbf{D}_w = \frac{1}{m}\sum_{j=1}^{m}\mathbf{D}_{w_j}. \tag{8}$$

This normalization restriction is for convenience; it simplifies the minimization procedure.

The minimization problem above can be formally stated as

$$\min_{F,U}\phi^{(ic)}(\mathbf{F},\mathbf{U}), \tag{9}$$

which may be carried out by first minimizing $\phi^{(ic)}$ with respect to $\mathbf{u}_j$ ($j = 1,\cdots,m$) conditional on $\mathbf{F}$, and then with respect to $\mathbf{F}$ subject to (7). This process is written as

$$\min_{F,U}\phi^{(ic)}(\mathbf{F},\mathbf{U}) = \min_{F}\min_{U|F}\phi^{(ic)}(\mathbf{F},\mathbf{U}) = \min_{F}\phi^{(ic)}(\mathbf{F},\hat{\mathbf{U}}) = \min_{F}\phi^{(ic)*}(\mathbf{F}), \tag{10}$$

where $\hat{\mathbf{U}}$ minimizes $\phi^{(ic)}(\mathbf{F},\mathbf{U})$ conditional on $\mathbf{F}$, and $\phi^{(ic)*}(\mathbf{F}) = \phi^{(ic)}(\mathbf{F},\hat{\mathbf{U}})$. The conditional minimum of $\phi^{(ic)}$ with respect to $\mathbf{U}$ (given $\mathbf{F}$) is obtained by

$$\hat{\mathbf{u}}_j' = (\mathbf{z}_j'\mathbf{D}_{w_j}\mathbf{z}_j)^{-1}\mathbf{z}_j'\mathbf{D}_{w_j}\mathbf{F} \qquad (j = 1,\cdots,m). \tag{11}$$

Putting this estimate of $\mathbf{u}_j'$ into (6), we obtain

$$\phi^{(ic)*}(\mathbf{F}) = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SS}(\mathbf{F}-\mathbf{P}_{z_j/D_{w_j}}\mathbf{F})_{D_{w_j}}, \tag{12}$$

where

$$\mathbf{P}_{z_j/D_{w_j}} = \mathbf{z}_j\hat{\mathbf{u}}_j' = \mathbf{z}_j(\mathbf{z}_j'\mathbf{D}_{w_j}\mathbf{z}_j)^{-1}\mathbf{z}_j'\mathbf{D}_{w_j}. \tag{13}$$

Criterion (12) can be further rewritten as

$$\phi^{(ic)*}(\mathbf{F}) = \mathrm{tr}(\mathbf{F}'\mathbf{D}_w\mathbf{F}) - \mathrm{tr}(\mathbf{F}'\mathbf{P}\mathbf{F}), \tag{14}$$

where $\mathbf{D}_w$ is as defined in (8), and

$$\mathbf{P} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{D}_{w_j}\mathbf{P}_{z_j/D_{w_j}}. \tag{15}$$

Since the first term in (14) is constant under (7), minimizing (14) with respect to $\mathbf{F}$ under (7) is equivalent to maximizing

$$\psi(\mathbf{F}) = \mathrm{tr}(\mathbf{F}'\mathbf{P}\mathbf{F}) \tag{16}$$

subject to the same normalization restriction. The maximum of (16) can be obtained by solving the generalized eigen-equation of the form

$$\mathbf{P}\mathbf{F} = \mathbf{D}_w\mathbf{F}\boldsymbol{\Delta}_r, \tag{17}$$

where $\mathbf{F}$ is the matrix of generalized eigenvectors corresponding to the r dominant generalized eigenvalues, and $\boldsymbol{\Delta}_r$ is the diagonal matrix of the r dominant generalized eigenvalues arranged in descending order. Once (17) is solved, $\mathbf{F}$ is scaled (multiplied by $n^{1/2}$) to satisfy (7).

It may be worthwhile noting that $\mathbf{F}$ obtained above is usually not columnwise centered, i.e., $\mathbf{F}'\mathbf{1}_n \neq \mathbf{0}_r$. To satisfy this condition, we need to include the term $-\mathbf{1}_n\boldsymbol{\mu}_j'$ for each j ($j = 1, \cdots, m$) in the optimization criterion (9) and estimate $\mathbf{F}$ in such a way that it is orthogonal to this term (Takane and Oshima-Takane, 2003). However, we do not pursue this possibility in this paper.

When there are no missing data, $\mathbf{D}_{w_j} = \mathbf{I}_n$ for all j, so that $\mathbf{D}_w = \mathbf{I}_n$. Then $\mathrm{GSVD}(\mathbf{D}_w^{-1}\mathbf{Z}^*\mathbf{S}^{-1})_{D_w,S}$ reduces to $\mathrm{GSVD}(\mathbf{Z}^*\mathbf{S}^{-1})_{I_n,S}$, which is essentially equivalent to PCA of the standardized data matrix $n^{1/2}\mathbf{Z}\mathbf{S}^{-1/2}$.

There is no unequivocal definition of the matrix of component loadings $\mathbf{A}$ in this formulation because the optimization criterion (6) is defined without this quantity. One natural choice, and the one we adopt in this paper, is $\mathbf{A} = \mathbf{C}_r\mathbf{D}_r/n^{1/2}$, where $\mathbf{C}_r$ and $\mathbf{D}_r$ are the portions of $\mathbf{C}$ and $\mathbf{D}$ pertaining to the r dominant (generalized) singular values. This is analogous to (3) for complete data.

One advantage of the MDP method is that the solution can be obtained non-iteratively. This sets us free from all kinds of problems associated with iterative procedures, e.g., non-convergence, choice of a stopping criterion, convergence to suboptimal solutions, etc. Non-iterative closed-form solutions also imply that solutions with different dimensionality are nested in the sense that lower dimensional solutions are merely subsets of higher dimensional solutions. This means that no *a priori* decision has to be made about the dimensionality of the solutions. The computation time is also relatively stable, dependent

mostly on the size of the data matrix. This is difficult to assess in iterative procedures because the number of iterations needed for convergence is difficult to know in advance.

2.2 The weighted low rank approximation (WLRA) method

We introduce the WLRA method by generalizing (5) to accommodate missing observations (Gabriel and Zamir 1979). Let $\mathbf{a}_j$ be the j-th column vector of $\mathbf{A}'$. Then, (5) can be rewritten as $\tau^{(c)}(\mathbf{F}, \mathbf{A}) = (1/m)\sum_{j=1}^{m} \mathrm{SS}(\mathbf{z}_j - \mathbf{F}\mathbf{a}_j)$. We generalize this criterion as

$$\tau^{(ic)}(\mathbf{F}, \mathbf{A}) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{SS}(\mathbf{z}_j - \mathbf{F}\mathbf{a}_j)_{D_{w_j}}, \tag{18}$$

where $\mathrm{SS}(\mathbf{y})_W = \mathbf{y}'\mathbf{W}\mathbf{y}$ for any column vector $\mathbf{y}$ and a symmetric nnd matrix $\mathbf{W}$. We minimize this criterion by alternately minimizing it with respect to $\mathbf{F}$ for fixed $\mathbf{A}$ and with respect to $\mathbf{A}$ for fixed $\mathbf{F}$.

It can be readily seen from (18) that the estimate of $\mathbf{a}_j$ that minimizes (18) for given $\mathbf{F}$ is obtained by

$$\hat{\mathbf{a}}_j = (\mathbf{F}'\mathbf{D}_{w_j}\mathbf{F})^{-1}\mathbf{F}'\mathbf{D}_{w_j}\mathbf{z}_j \qquad (j = 1, \cdots, m). \tag{19}$$

To obtain the estimate of $\mathbf{F}$ for given $\mathbf{A}$, we rewrite (18) as follows: Let $\mathbf{z}'_{(i)}$ denote the i-th row vector of $\mathbf{Z}$, and let $\mathbf{f}'_{(i)}$ denote the i-th row vector of $\mathbf{F}$ $(i = 1, \cdots, n)$. Let $\mathbf{D}_{w_{(i)}}$ denote the diagonal matrix whose kth diagonal element is unity if the k-th element of $\mathbf{z}'_{(i)}$ is observed, and zero if it is not observed. Then,

$$\tau^{(ic)}(\mathbf{F}, \mathbf{A}) = \frac{1}{m}\sum_{i=1}^{n} \mathrm{SS}(\mathbf{z}'_{(i)} - \mathbf{f}'_{(i)}\mathbf{A}')_{D_{w_{(i)}}}, \tag{20}$$

where $\mathrm{SS}(\mathbf{y}')_W = \mathrm{tr}(\mathbf{y}'\mathbf{W}\mathbf{y})$ for any row vector $\mathbf{y}'$ and a symmetric nnd matrix $\mathbf{W}$. It can be observed that the conditional minimum of (20) with respect to $\mathbf{f}^{(i)}$ for given $\mathbf{A}$ is obtained by

$$\hat{\mathbf{f}}'_{(i)} = \mathbf{z}'_{(i)}\mathbf{D}_{w_{(i)}}\mathbf{A}(\mathbf{A}'\mathbf{D}_{w_{(i)}}\mathbf{A})^{-1} \qquad (i = 1, \cdots, n). \tag{21}$$

As has been noted above, we apply (19) and (21) alternately to update $\mathbf{A}$ and $\mathbf{F}$ until convergence is reached. We may stop the iteration as soon as the change in the value of $\tau^{(ic)}$ from one iteration to the next gets smaller than a certain value, e.g., 10^{-10}, as we adopted in this paper. The above algorithm is called the criss-cross algorithm (Gabriel and Zamir 1979). It is a special kind of alternating least squares (ALS) algorithms, and consequently it is monotonically convergent. We need an initial estimate of $\mathbf{A}$ to start the algorithm. We may randomly generate an initial $\mathbf{A}$, or alternatively we may use $\mathbf{A}$ obtained by SVD of $\mathbf{Z}^*$. The latter tends to lead to faster convergence. (In the simulation studies to be reported in the next section, we used $\mathbf{A}$ obtained

from original complete data as an initial estimate of $\mathbf{A}$.) This algorithm is simply referred to as the iterative algorithm (IA) in Folch-Fortuny et al. (2015; Walczak and Massart 2001).

After convergence is reached, we re-scale $\mathbf{F}$ to satisfy the normalization restriction (7). Specifically, let $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{A}}$ denote the estimates of $\mathbf{F}$ and $\mathbf{A}$ before re-scaling. Let $\tilde{\mathbf{F}} = \tilde{\mathbf{Q}}\tilde{\mathbf{G}}\tilde{\mathbf{R}}'$ represent the (compact) SVD of $\tilde{\mathbf{F}}$. Then, the re-scaled estimates of $\mathbf{F}$ and $\mathbf{A}$ are obtained by $\mathbf{F} = n^{1/2}\tilde{\mathbf{Q}}$ and $\mathbf{A} = \tilde{\mathbf{A}}\tilde{\mathbf{R}}\tilde{\mathbf{G}}/n^{1/2}$. Note that $\tilde{\mathbf{F}}\tilde{\mathbf{A}}' = \mathbf{F}\mathbf{A}'$, which justifies this re-scaling procedure. It should be noticed that $\mathbf{F}$ obtained this way is not mean-centered, i.e., $\mathbf{F}'\mathbf{1}_n \neq \mathbf{0}_r$, as in the MDP method. There is, in fact, a way to isolate means from $\mathbf{F}$ and force it to be mean centered. We do not follow this procedure in this paper in line with the MDP method in which $\mathbf{F}$ is left not mean-centered.

The WLRA method presented above specifically designed to handle missing data was originally proposed by Gabriel and Zamir (1979). See also Grung and Manne (1998). Their procedure has been subsequently generalized to a method for obtaining weighted low rank approximations of data matrices under very flexible weighting schemes by Wentzell, Andrews, Hamilton, Faber, and Kowalski (1997), which subsumes the missing data case as a special case.

The WLRA method explicitly aims to obtain a matrix of lower rank which best approximates a data matrix. The solution, however, is iterative, and consequently possibilities of non-convergence or convergence to suboptimal solutions cannot be ruled out. Derived solutions are not nested, and so the solutions in different dimensionality must be obtained separately. This implies that the dimensionality must be prescribed in advance.

It is instructive to note that there is an interesting variant of the above algorithm. Let $\tilde{\mathbf{Z}}$ denote an optimally scaled data matrix, by which we mean that the elements of $\tilde{\mathbf{Z}}$ corresponding to observed data are copied from the corresponding elements of the data matrix $\mathbf{Z}$, while those corresponding to the missing data are copied from the corresponding elements in the matrix of best approximation (i.e., $\mathbf{F}\mathbf{A}'$). Then $\tau^{(ic)}$ above can be restated as:

$$\tau^{(ic)}(\mathbf{F}, \mathbf{A}) = \mathrm{SS}(\tilde{\mathbf{Z}} - \mathbf{F}\mathbf{A}'). \tag{22}$$

Note that missing data are always fitted perfectly in this set-up, which is equivalent to zeroing the misfit of missing data by zero weights. The minimization of this criterion with respect to $\mathbf{F}$ and $\mathbf{A}$ for fixed $\tilde{\mathbf{Z}}$ is a complete data problem that can be solved in a number of different ways. The easiest way (if not the most efficient way) is via SVD of $\tilde{\mathbf{Z}}$. See Eqs. (1), (2), and (3), where $\mathbf{Z}$ in (1) is replaced by the current $\tilde{\mathbf{Z}}$. Once the new estimates of $\mathbf{F}$ and $\mathbf{A}$ are obtained, $\mathbf{F}\mathbf{A}'$ is calculated and $\tilde{\mathbf{Z}}$ is updated by copying the relevant elements of $\mathbf{F}\mathbf{A}'$. We may alternate the estimation of $\mathbf{F}$ and $\mathbf{A}$ and that of $\tilde{\mathbf{Z}}$ until convergence is reached. This is also an ALS algorithm and is known to converge to essentially identical points to the algorithm given earlier. This approach is called an optimal scaling approach to missing data, which turns out to be a special case of the WLRA method (Grung and Manne 1998; Kiers 1997).

2.3 The regularized PCA (RPCA) method

It has been pointed out that the WLRA method described above sometimes overfits missing data (Grung and Manne 1998; van Ginkel et al. 2014). This means that PCA solutions are predominantly influenced by imputed values for missing data. To overcome this difficulty, Josse et al. (2009; Josse and Husson 2012) proposed a so-called regularized PCA (RPCA) method. This method works in a manner similar to the usual (non-regularized) WLRA method, except that the following formulas are used to calculate $\mathbf{F}$ and $\mathbf{A}$, instead of (2) and (3). Let $\mathbf{B}_r$, $\mathbf{D}_r$, and $\mathbf{C}_r$ be matrices analogous to those used in (2) and (3) obtained from the current imputed data matrix. Then

$$\mathbf{F}^{(R)} = n^{1/2}\mathbf{B}_r(\mathbf{D}_r^2 - \hat{\sigma}^2\mathbf{I}_r)^{1/2}\mathbf{D}_r^{-1}, \tag{23}$$

and

$$\mathbf{A}^{(R)} = n^{-1/2}\mathbf{C}_r(\mathbf{D}_r^2 - \hat{\sigma}^2\mathbf{I}_r)^{1/2}, \tag{24}$$

where

$$\hat{\sigma}^2 = \mathrm{tr}(\mathbf{D}_{m-r}^2)/(m-r). \tag{25}$$

Here, $\mathbf{D}_{m-r}$ is the portion of $\mathbf{D}$ (the diagonal matrix of the entire set of singular values) corresponding to the $m-r$ smallest singular values. The formulas lead to

$$\mathbf{F}^{(R)}\mathbf{A}^{(R)'} = \mathbf{B}_r(\mathbf{D}_r - \hat{\sigma}^2\mathbf{D}_r^{-1})\mathbf{C}_r', \tag{26}$$

from which imputed values for missing data (model predictions corresponding to missing data) are copied into the current data matrix. The SVD is then reapplied to the updated data matrix, and $\mathbf{F}^{(R)}$ and $\mathbf{A}^{(R)}$ are recalculated. This process is repeated until no substantial change occurs in imputed values from one iteration to the next.

The above formulas for regularized component-loading and score matrices have been derived from the probabilistic PCA model (Tipping and Bishop 1999), in which not only errors but also component scores are assumed to be random vectors. Let $\mathbf{f}$ denote the random (column) vector of component scores representing a row of $\mathbf{F}$, and let $\mathbf{e}$ denote the random vector of measurement errors. Then the probabilistic PCA model can be written as

$$\mathbf{z} = \mathbf{A}\mathbf{f} + \mathbf{e}, \tag{27}$$

where $\mathbf{z}$ is a random vector of observed variables. It is further assumed that $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, and $\mathbf{f}$ and $\mathbf{e}$ are statistically independent from each other. It follows that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\mathbf{A}' + \sigma^2\mathbf{I}_m)$. (This is like a random-effect common factor analysis model with constant unique variances.) A maximum likelihood estimate of $\mathbf{A}$ under this model is obtained by (24) (Tipping and Bishop 1999). A vector of component scores $\mathbf{f}$, which is a random vector, cannot be estimated in the usual sense, but it is customary to take the expectation of posterior density of $\mathbf{f}$ given $\mathbf{z}$ as its predictor, which is given by

$$\mathrm{E}[\mathbf{f}_i|\mathbf{z}_i] = n^{1/2}\mathbf{A}'(\mathbf{A}\mathbf{A}' + \sigma^2\mathbf{I}_m)^{-1}\mathbf{z}_i = n^{1/2}(\mathbf{A}'\mathbf{A} + \sigma^2\mathbf{I}_r)^{-1}\mathbf{A}'\mathbf{z}_i, \tag{28}$$

where $\mathbf{f}_i'$ and $\mathbf{z}_i'$ are the ith row vectors of $\mathbf{F}$ and $\mathbf{Z}$. The entire matrix of $\mathbf{F}$ is obtained by (23) since $\mathbf{A}(\mathbf{A}'\mathbf{A} + \sigma^2\mathbf{I}_r)^{-1} = (\mathbf{D}_r^2 - \sigma^2\mathbf{I}_r)^{1/2}\mathbf{D}_r^{-1}$. The $\hat{\sigma}^2$ in (23) and (24) is an estimate of σ^2. It may be noted that this estimate of $\mathbf{F}$ is equivalent to the regression estimate of factor scores in common factor analysis under the constant unique variance assumption (McDonald and Burr 1967).

2.4 The trimmed scores regression (TSR) method

This method uses an algorithm similar to the one described in the prvious section. It alternates between two stages: One obtains parameters in PCA (i.e., $\mathbf{F}$ and $\mathbf{A}$) for given data (with imputed values for missing data), and the other updates imputed values given the parameters. The difference is that while in the WLRA method, imputed values for missing data are obtained by the corresponding model values (i.e., elements of $\mathbf{FA}'$), in the TSR method, they are obtained by regression (Folch-Fortuny et al. 2015). Let $\tilde{\mathbf{z}}_{i,t}'$ denote the ith row of the data matrix $\tilde{\mathbf{Z}}_t$ in iteration t. It is convenient to arrange the elements of $\tilde{\mathbf{z}}_{i,t}'$ in such way that the elements corresponding to missing values in the original data all come at the beginning, followed by those corresponding to observed values. This is expressed as $\tilde{\mathbf{z}}_{i,t}' = [\tilde{\mathbf{z}}_{i,t}^{(M)'}, \tilde{\mathbf{z}}_{i,t}^{(O)'}]$, where the parenthesized superscripts $^{(M)}$ and $^{(O)}$ stand for missing data and observed data parts, respectively. The part corresponding to the missing data is updated in each iteration, while the observed data part remains constant. We also rearrange columns of $\tilde{\mathbf{Z}}_{i,t}$ conformably to the rearrangement of its ith row vector. This is written as $\tilde{\mathbf{Z}}_{i,t} = [\tilde{\mathbf{Z}}_{i,t}^{(M)}, \tilde{\mathbf{Z}}_{i,t}^{(O)}]$. Note that this rearrangement is induced by the missing data pattern in the ith row of the original data matrix, which is the reason why the subscript i is put on $\tilde{\mathbf{Z}}_{i,t}$ (i.e., it depends on i). Define

$$\mathbf{S}_{i,t} = \begin{bmatrix} \mathbf{S}_{i,t}^{(MM)} & \mathbf{S}_{i,t}^{(MO)} \\ \mathbf{S}_{i,t)}^{(OM)} & \mathbf{S}_{i,t}^{(OO)} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{Z}}_{i,t}^{(M)'}\tilde{\mathbf{Z}}_{i,t}^{(M)} & \tilde{\mathbf{Z}}_{i,t}^{(M)'}\tilde{\mathbf{Z}}_{i,t}^{(O)} \\ \tilde{\mathbf{Z}}_{i,t}^{(O)'}\tilde{\mathbf{Z}}_{i,t}^{(M)} & \tilde{\mathbf{Z}}_{i,t}^{(O)'}\tilde{\mathbf{Z}}_{i,t}^{(O)} \end{bmatrix}. \tag{29}$$

Finally, let $\mathbf{A}_{i,t}' = [\mathbf{A}_{i,t}^{(M)'}, \mathbf{A}_{i,t}^{(O)'}]$ denote the estimate of $\mathbf{A}'$ at iteration t whose columns are arranged similarly to $\tilde{\mathbf{Z}}_{i,t}$. Then, the missing data part of the ith row of $\tilde{\mathbf{Z}}_{i,t}$ is updated (for the next iteration) by

$$\tilde{\mathbf{z}}_{i,t+1}^{(M)'} = \tilde{\mathbf{z}}_{i,t}^{(O)'}\mathbf{A}_{i,t}^{(O)}(\mathbf{A}_{i,t}^{(O)'}\mathbf{S}_{i,t}^{(OO)}\mathbf{A}_{i,t}^{(O)})^{-}\mathbf{A}_{i,t}^{(O)'}\mathbf{S}_{i,t}^{(OM)}, \tag{30}$$

where $^{-}$ indicates a g-inverse. This formula is obtained as the ith row vector of the best prediction of $\tilde{\mathbf{Z}}_{i,t}^{(M)}$ from $\tilde{\mathbf{Z}}_{i,t}^{(O)}\mathbf{A}_{i,t}^{(O)}$ as predictors. (The word "trimmed" in the name stems from the fact that $\tilde{\mathbf{Z}}_{i,t}^{(O)}\mathbf{A}_{i,t}^{(O)}$ is used as predictors, not $\tilde{\mathbf{Z}}_{i,t}^{(O)}$ per se, that is, $\tilde{\mathbf{Z}}_{i,t}^{(O)}$ is trimmed by $\mathbf{A}_{i,t}^{(O)}$.) This formula is applied for all rows with missing data in each iteration with $\mathbf{S}_{i,t}$ and $\mathbf{A}_{i,t}$ redefined for each i. The whole iteration is repeated until the change in successive

updates of $\tilde{\mathbf{z}}_{i,t+1}^{(M)'}$ is smaller than a certain threshold. We used a MATLAB routine called pcambtsr.m in the Missing Data Imputation (MDI) Toolbox developed by Folch-Fortuny, Arteaga, and Ferrer (2016). We also used the default convergence criterion of 10^{-10}.

2.5 The data augmentation (DA) method

In contrast to the previous methods, the data augmentation (DA) method (Schafer, 1997) imputes more than one value (say K values) for each missing observation. The method consists of two major loops, an outer loop and an inner loop. The outer loop generates K sets of imputed values and parameters in the PCA model presumed to have generated the data. In the present case, model parameters in PCA comprise a mean vector and a covariance matrix of a multivariate normal distribution. The inner loop, on the other hand, generates each of the K sets of imputed values and a mean vector and a covariance matrix. In the inner loop, the mean vector $\mathbf{m}$ and the covariance matrix $\mathbf{S}$ are first initialized, and then the following two steps are iterated for a fixed number of times:

1. Imputation Step: Draw imputed values $\tilde{\mathbf{Z}}_t^{(M)}$ from the distribution of missing data given the observed data $\tilde{\mathbf{Z}}_t^{(O)}$, the mean vector $\mathbf{m}_t$, and the covariance matrix $\mathbf{S}_t$ in iteration t.

2. Posterior Step: Draw the mean vector and the covariance matrix for the next iteration (i.e., $\mathbf{m}_{t+1}$ and $\mathbf{S}_{t+1}$) from their posterior distributions given $\tilde{\mathbf{Z}}_t^{(M)}$ and $\tilde{\mathbf{Z}}_t^{(O)}$.

The two steps are repeatedly applied until the posterior distributions of $\mathbf{m}$ and $\mathbf{S}$ stabilize. In the MATLAB program we used (see below), the above two steps are applied for a fixed number of times (CL) presumed to be large enough for stabilization. The procedure of alternately simulating missing data and model parameters as above forms a Markov chain that eventually converges in distribution (Schafer 1997; Tanner and Wong 1887). We used the MATLAB routine "pcambda.m" provided in the MDI Toolbox (Folch-Fortuny et al. 2016) for our computation with the default values of $K = 10$ and $CL = 100$.

Once K sets of mean vectors and covariance matrices are obtained, a single set of imputed data are generated for each set of mean vector and covariance matrix by applying the Imputation Step once more. PCA is then applied to each completed data set to derive a loading matrix $\mathbf{A}$ and a score matrix $\mathbf{F}$. This is repeated K times, and K PCA solutions are obtained. A single component loading matrix is then derived by applying a multiple-set Procrustes rotation procedure similar to the one used in Van Ginkel and Kroonenberg (2014).

Note that there may be other implementations of the DA method (e.g., Schaffer 1997) that may be more efficient than the one implemented here, so

that when we say the DA method in this paper, it refers to the particular implementation of the DA method as described above.

3 Empirical studies

It is of interest to compare the five methods described above under various conditions because to the best of our knowledge, there have been no studies that compared all of them simultaneously. In what follows, we first provide the general set-up of the simulation studies, and then specific details.

We first select several real data sets of varying sizes published in the the literature. These data are initially complete. For each data set, we generate many (100) incomplete data sets by randomly and non-randomly creating missing data in varying proportions (10, 20, and 30%). We analyze them by the five methods with varying numbers of components. We compare their parameter recovery capability as functions of the missing data proportions (censor rate), the number of extracted components (dimensionality), and the degree of non-randomness in censoring processes. For convenience, the censor rate of 10% is called a "small," 20% "medium," and 30% "large" censor rate. Sections 3.1 through 3.3 address missing completely at random (MCAR) situations, while Section 3.4 deals with missing not completely at random (MNCAR) situations.

Let $\mathbf{Z}$ represent the original standardized data matrix. By applying PCA to $\mathbf{Z}$, we obtain $\mathbf{F}_r$ and $\mathbf{A}_r$ for specific dimensionality r. We may also calculate $\mathbf{Z}_r = \mathbf{F}_r\mathbf{A}_r'$, which gives the best rank r approximation to $\mathbf{Z}$. Let $\mathbf{Z}^{(q)}$ represent the qth censored data ($q = 1, \cdots, 100$). By applying either one of the five missing data methods described in the previous section, we obtain $\mathbf{F}_r^{(q)}$ and $\mathbf{A}_r^{(q)}$, from which we may also calculate $\mathbf{Z}_r^{(q)}$. In this paper, the goodness of parameter recovery is assessed by the congruence coefficient (Tucker 1951) between $\mathbf{A}_r$ and $\mathbf{A}_r^{(q)}$. Obviously, the same thing can be done between $\mathbf{F}_r$ and $\mathbf{F}_r^{(q)}$, and between $\mathbf{Z}_r$ and $\mathbf{Z}_r^{(q)}$. In this paper, however, we focus on the recovery of the loading matrix, since preliminary analyses indicate that patterns of recovery are very similar among them. The congruence coefficient is defined as

$$c(\boldsymbol{\theta}_o, \boldsymbol{\theta}_q) = \frac{\boldsymbol{\theta}_o'\boldsymbol{\theta}_q}{\sqrt{\boldsymbol{\theta}_o'\boldsymbol{\theta}_o \cdot \boldsymbol{\theta}_q'\boldsymbol{\theta}_q}}, \tag{31}$$

where $\boldsymbol{\theta}_o = \text{vec}(\mathbf{A}_r)$, and $\boldsymbol{\theta}_q = \text{vec}(\mathbf{A}_r^{(q)})$. The closer is the value of this coefficient to unity, the better is the recovery. For each of the five methods, and for each censor rate and dimensionality of solutions, mean and standard deviation of c is calculated over repeated censoring ($q = 1, \cdots, 100$). In the sequel, a mean congruence coefficient of .99 or above is termed "excellent" recovery, that of .95 or above is called "good" recovery, and that of .90 or above is called "acceptable" recovery.

3.1 The food and cancer data under the MCAR condition

The first data set we analyze is a small data set compiled by Segi (1979). There are six variables in total, of which four variables related to food (1. Average daily calories, 2. Meat supply, 3. Supply of milk products, and 4. Alcohol consumption) were initially gathered by FAO (Food and Agriculture Organization), while the remaining two related to cancer mortality rates (1. Large intestine, and 2. Rectum) were collected by WHO (World Health Organization). The data were gathered in 47 countries in the world. The original data set has been provided in Table 1.3 of Takane (2013). For the purpose of this study, the initially complete data were censored completely at random (MCAR) in prescribed proportions (10%, 20%, and 30%). The MCAR means that every element in the original data has equal chance of being censored.

PCA was first applied to the original complete data, which revealed that the first four components accounted for 70.8%, 14.1%, 6.2%, and 5.3% of the total variation in the standardized data. It seemed that there were two empirically significant components, one strong and the other relatively weak. It was decided to examine the number of components from 1 to 3.

Table 1 summarizes the main results. The first column (labeled "dim.") of the table indicates the number of extracted components (dimensionality). The second column (labeled "p") indicates the censoring rate. The next five columns show the mean and standard deviation (in parentheses) of the congruence coefficient for component loadings obtained by the five methods. Note that the table presents only the results for dimensionality 2 and 3. The results for dimensionality 1 are omitted because all five methods provide the mean congruence coefficients of nearly one with nearly zero standard deviations across all censor rates. In general the recovery rate is excellent for small numbers of components and low censoring rates. The recovery rate deteriorates, as the dimensionality and the censor rate increase. However, the rate of deterioration in recovery is not uniform across the five methods. The RPCA method, the DA method, and the TSR method maintain excellent to good recovery rates across all conditions. The recovery by the MDP method is slightly but consistently worse than these three methods, although it still maintains an acceptable level across all conditions. The WLRA works almost as well as the other methods when the dimensionality is small to moderate, but its recovery rate falls below the acceptable level when both the dimensionality and the censor rate are high.

Note that for dimensionality greater than one, the recovery rate reported in Table 1 reflects an average recovery rate over multiple components, but there may be variations in recovery across the components. This conjecture was indeed supported by Table 2 reporting componentwise recovery rates. There is a clear tendency that more dominant components are better recovered than less dominant components, although there are also some variations in this tendency across different methods. The decline in recovery is fastest in MDP method, and slowest in the WLRA method, while the remaining three methods fall between them, which are also the best methods overall, achieving the

Table 1 Recovery of component loadings for the food-cancer data: Means and standard deviations (in parentheses) of the congruence coefficient

dim	p	MDP	WLRA	RPCA	TSR	DA
	10%	1.0	1.0	1.0	1.0	1.0
		(.00)	(.00)	(.00)	(.00)	(.00)
2	20%	.98	.99	1.0	1.0	1.0
		(.04)	(.04)	(.01)	(.00)	(.00)
	30%	.97	.90	.99	.98	.99
		(.04)	(.10)	(.02)	(.02)	(.02)
	10%	.98	.99	.99	1.0	.99
		(.02)	(.01)	(.01)	(.00)	(.01)
3	20%	.95	.91	.99	.99	.98
		(.04)	(.15)	(.01)	(.01)	(.02)
	30%	.92	.85	.97	.97	.97
		(.04)	(.18)	(.02)	(.03)	(.04)

acceptable level of recovery for the first two components across all conditions. Note, however, that no methods achieved the acceptable level recovery for the third component under any censor rates examined.

Table 2 Componentwise recovery of loadings for the food-cancer data: Mean congruence coefficients as functions of the censor rate and the method

	No. of dimensions	2		3		
p	Method\Comp.	1	2	1	2	3
	MDP	1.0	.98	1.0	.98	.79
	WLRA	1.0	1.0	1.0	.98	.87
10%	RPCA	1.0	.99	1.0	.99	.87
	TSR	1.0	.99	1.0	.99	.89
	DA	1.0	.99	1.0	.99	.88
	MDP	1.0	.95	1.0	.94	.66
	WLRA	1.0	.99	.96	.87	.86
20%	RPCA	1.0	.98	1.0	.97	.78
	TSR	1.0	.98	1.0	.98	.81
	DA	1.0	.98	1.0	.98	.80
	MDP	1.0	.88	1.0	.90	.56
	WLRA	.93	.92	.92	.81	.82
30%	RPCA	1.0	.97	1.0	.96	.73
	TSR	1.0	.97	1.0	.94	.70
	DA	1.0	.95	1.0	.96	.75

3.2 The organizational identity data under the MCAR condition

The second data set we analyze is a much larger data set. It is part of the survey data used in Bergami and Bagozzi (2000), consisting of a sample of 305 employees (male = 157 and female = 148) from the electronics division of a large conglomerate in South Korea. From the original data, Hwang and Takane

(2014, Section 3.3.1) used twenty one variables, which we also use in the present study. They fitted a structural equation model to this data set with four latent variables, named organizational prestige (OP), organizational identity (OI), affective commitment - joy (ACJ), and affective commitment - love (ACL). They assumed that eight variables are associated with OP, six variables with OI, four variables with ACJ, and the remaining three variables with ACL. An example of a variable that represents OP is: My relatives and people close or important to me believe that [Company X] is a highly respected company. An example of an indicator variable for OI is: When someone criticizes [Company X] it feels like a personal insult, that for ACJ is: I would be very happy to spend the rest of my career with [Company X], and that for ACL is: I do not feel like part of a family at [Company X]. The complete list of the twenty one variables is presented in Table 3.7 of Hwang and Takane (2014, p. 113). Subjects were asked to indicate how much they agreed or disagreed to the statements using 5-point rating scales: 1. strongly disagree, 2. disagree, 3. neither agree nor disagree, 4. agree, and 5. strongly agree.

We are tempted to assume that there are four distinct components corresponding to the four latent variables noted above. However, they are all highly correlated with each other. PCA was applied to the original complete data, which revealed that the first six components accounted for 35.0%, 14.9%. 6.9%, 5.3%, 4.5%, and 3.7%, respectively, totaling 70.2% of the total variation in the standardized data. As it seems, the first two components are rather strong, while the remaining four are relatively weak. It was decided to examine the dimensionality between 1 and 6 inclusive.

The design of the simulation study remains essentially the same as in the previous study. The major results are reported in Table 3. Results on one and two components are omitted in this table since all five methods achieved near perfect recovery (the mean congruence coefficients of 1.0 with near zero standard errors). The general pattern of the results is similar to Study 1. The recovery of component loadings is good to excellent across all five methods and censor rates up to three-component solutions. In four-component solutions, the WLRA method falls below the acceptable level for highly censored data, while the other four methods maintain good recovery across all censor rates. In five-component solutions, the WLRA method falls below the good recovery level even for the small censor rate and below the acceptable level for medium-sized and highly censored data. The MDP method also falls below good recovery for highly censored data, while the other three methods maintain the good recovery level. In six-component solutions, the WLRA method fails to achieve the acceptable level for all censor rates, and the MDP method for highly censored data, while the other three methods maintain the acceptable level for all censor rates. Among the three best performing methods, the RPCA method seems to have a slight edge over the other two methods, although the difference is minor.

Table 3 reported only average recovery rates over components for multi-component solutions. As in the previous study, the recovery rate could vary over the components. We therefore evaluated componentwise recovery rates as

Table 3 Recovery of component loadings for the organizational identity data: Means and standard deviations (in parentheses) of the congruence coefficient

dim	p	MDP	WLRA	RPCA	TSR	DA
3	10%	1.0	1.0	1.0	1.0	1.0
		(.00)	(.00)	(.00)	(.00)	(.00)
	20%	.99	.99	.99	1.0	1.0
		(.00)	(.01)	(.00)	(.00)	(.00)
	30%	.98	.98	.99	.99	.99
		(.02)	(.02)	(.01)	(.00)	(.00)
4	10%	.99	.99	.99	1.0	1.0
		(.00)	(.01)	(.00)	(.00)	(.00)
	20%	.98	.93	.99	.99	.99
		(.01)	(.15)	(.01)	(.01)	(.01)
	30%	.96	.86	.98	.98	.97
		(.03)	(.19)	(.02)	(.01)	(.02)
5	10%	.99	.94	.99	.99	.99
		(.01)	(.14)	(.00)	(.01)	(.00)
	20%	.96	.79	.98	.98	.98
		(.02)	(.22)	(.01)	(.01)	(.01)
	30%	.94	.58	.97	.96	.96
		(.02)	(.25)	(.02)	(.02)	(.02)
6	10%	.97	.84	.98	.99	.99
		(.02)	(.23)	(.01)	(.01)	(.01)
	20%	.94	.58	.97	.97	.97
		(.02)	(.26)	(.02)	(.01)	(.02)
	30%	.91	.38	.95	.95	.94
		(.02)	(.18)	(.02)	(.02)	(.02)

in Table 2. A similar tendency to the previous study was observed. For details, see Table A1 in Online Resource. The rate of decline in parameter recovery over the components is fastest with the MDP method, and slowest with the WLRA method. The three remaining methods (RPCA, TSR, and DA) achieved the acceptable level of recovery up to the fifth component for small censoring, up to the fourth component for medium-sized censoring, and up to the third component for highly censored data, for any dimensional solutions.

3.3 Mezzich's data under the MCAR condition

So far, all of our example data sets had more rows than columns. While in a majority of situations in which PCA is applied, this is the case, what happens if it is not true? Our third example addresses this question. To this end, we use Mezzich's (1978) data collected from eleven psychiatrists rating four archetypal psychiatric patients, 1) manic depressive - depressed (MDD), 2) manic depressive - manic (MDM), 3) simple schizophrenic (SSP), and 4) paranoid schizophrenic (PSP), using seventeen Brief Psychiatric Rating Scales (BPRS) by Overall and Gorham (1962). Each of the seventeen scales has seven ordered categories (ranging from 0 indicating "Does not apply at all" to 6 indicating "Applies very well"). Examples of the seventeen scales are: Somatic concern,

Anxiety, Emotional withdrawal, etc. The entire data set consisting of 44 rows and 17 columns is given in Table 1.1 in Takane (2013).

In Takane (2013), PCA was applied to rowwise centered data to emphasize the contrasts among the scales. This was like analyzing the transposed data. The transposition of the data creates an example in which the number of columns (44) exceeds the number of rows (17). Comparing the performance of the methods under this condition is interesting, to see if the tendency we have observed above in the standard set-up (where the number of rows is larger than that of columns) remains valid. Note, however, that only four methods (the MDP, WLRA, RPCA, and TSR methods) are up for comparison in this study. The DA method, requiring invertible covariance matrices, is not feasible when the number of rows is larger than that of columns.

PCA of the original (complete) data set indicated that the first six components explained 48.7%, 17.8%, 10.7%, 7.5%, 3.7%, and 2.5% of the total variation in the standardized data. It was decided to study 1 to 6-component solutions. Theoretically, three dimensions are sufficient to discriminate four groups of patients, and so six components seem to be an over-extraction, although there may be substantial individual differences among the psychiatrists. Table 4 provides a summary of the results, which remain similar to those reported earlier despite the difference in the data profile. The parameter recovery is good to excellent up to two-component solutions for all censor rates and methods. In three-component solution, some methods begin to suffer for highly censored data, although the recovery rate is still acceptable. A problem starts in four-component solutions, in which the WLRA, and TSR methods fail to reach the acceptable level of recovery for highly censored data, while the remaining two (MDP and RPCA) maintain the acceptable level. In five-component solutions, results are similar to those in four-component solutions, except that the WLRA method falls below the the acceptable level earlier (at the moderate censor level). In six-component solutions, all four methods fail to reach the acceptable level for highly censored data. Overall, the RPCA method worked best. The parameter recovery tends to be somewhat lower across all conditions in this data set than in the previous data sets. This is partly because the parameter recovery is measured in terms of component loadings. The data having a larger number of columns than rows tend to have better recovery in component scores.

Componentwise recovery rates indicated a similar tendency as in the previous studies, except that the RPCA method seemed to work clearly better than the TSR method. The RPCA achieved the acceptable level of recovery up to the fifth component for small censoring, up to the fourth component for medium-size censoring, but only up to the second component for highly censored data, regardless of dimensionality. See Table A2 in Online Resource for more detail.

Table 4 Recovery of component loadings for the Mezzich's data: Means and standard deviations (in parentheses) of the congruence coefficient

dim	p	MDP	WLRA	RPCA	TSR	dim	p	MDP	WLRA	RPCA	TSR
	10%	1.0	1.0	1.0	1.0		10%	.98	.99	.99	.99
		(.00)	(.00)	(.00)	(.00)			(.01)	(.01)	(.00)	(.01)
1	20%	.99	.99	.99	.99	4	20%	.96	.95	.98	.96
		(.00)	(.00)	(.00)	(.00)			(.02)	(.10)	(.01)	(.02)
	30%	.99	.99	.98	.98		30%	.93	.82	.94	.88
		(.00)	(.00)	(.01)	(.01)			(.05)	(.21)	(.02)	(.11)
	10%	.99	1.0	.99	.99		10%	.98	.97	.99	.98
		(.00)	(.00)	(.00)	(.00)			(.01)	(.08)	(.00)	(.01)
2	20%	.98	.99	.99	.98	5	20%	.94	.83	.97	.92
		(.01)	(.01)	(.01)	(.01)			(.03)	(.23)	(.01)	(.07)
	30%	.97	.98	.97	.96		30%	.91	.56	.92	.75
		(.01)	(.01)	(.02)	(.04)			(.03)	(.26)	(.04)	(.17)
	10%	.99	.99	.99	.99		10%	.97	.94	.98	.97
		(.00)	(.01)	(.00)	(.01)			(.02)	(.10)	(.01)	(.03)
3	20%	.98	.99	.98	.97	6	20%	.93	.70	.95	.83
		(.01)	(.01)	(.01)	(.01)			(.02)	(.28)	(.03)	(.15)
	30%	.96	.97	.95	.93		30%	.88	.37	.89	.57
		(.02)	(.02)	(.02)	(.06)			(.03)	(.21)	(.05)	(.22)

3.4 Recovery under the MNCAR conditions

So far, we examined the performance of various methods of PCA under the condition of missing completely at random (MCAR). In many practical situations, however, missing data arise in not completely at random (MNCAR) fashions. In this section, we investigate the effects of non-randomness in generating missing data on parameter recovery. There are infinitely many ways in which non-randomness occurs in missing data mechanisms (Josse, Timmerman, Kiers, 2013). Due to space limitations, we consider only limited sets of non-randomness conditions in this paper. Non-randomness in missing data mechanisms mean that the probability of missing data in certain variables (which we call "target" variables) depends on the values of the target variables themselves or of some other variables (which we call "agent" variables). More specifically, we assume that observations on a target variable are rendered missing whenever its agent variable takes one of the q largest values, where q is determined in such a way that the overall censor rate is equal to some prescribed values.

The target variables in each data set are chosen somewhat arbitrarily. For the food-cancer data, the target variables are Variables 1 and 2, and the value of q is set to 14 or 15 (approximately equal to $47 \times 6 \times .1 \div 2$) for 10% censoring, 28 or 29 for 20% censoring, and 42 or 43 for 30% censoring on each of the 2 target variables. For the organizational identity data, the target variables are Variables 1 to 4, 9 to 11, 15 to 16, and 19, and the value of q is set to 64 or 65 (approximately equal to $305 \times 21 \times .1 \div 10$) for 10% censoring, 128 or 129 for 20% censoring, and 192 or 193 for 30% censoring on each of the 10 target variables. For Mezzich's data, the target variables are Variables 1 to 5,

12 to 16, 23 to 27, and 34 to 38. For this data set, a 20% censoring created a target variable whose observed values are all the same. To avoid this to happen, the censoring rate was reduced to one half (i.e., 5%, 10% and, 15%) of the original proportions. The value of q is set to 1 or 2 (approximately equal to $17 \times 44 \times .05 \div 20$) for 5% censoring, 3 or 4 for 10% censoring, and 5 or 6 for 15% censoring on each of the 20 target variables.

We consider the following three scenarios for the choice of the agent variables:

MNCAR 1: The agent variables are randomly generated outside variables which are only randomly correlated with the target variables.

MNCAR 2: The agent variables are variables in the data set other than the target variables, which happen to be fairly highly correlated with the target variables.

MNCAR 3: The agent variables are the target variables themselves.

The correlations between the target and agent variables are lowest in MNCAR 1, largest in NMCAR 3, and in-between in MNCAR 2. We expect that MNCAR 1 is closest to MCAR, MNCAR 3 furthest from MCAR, and MNCAR 2 between the two. MNCAR 1 is most akin to MCAR with a primary difference being that in the former, missing data are concentrated in a few variables, while in the latter, they are distributed evenly over all variables. MNCAR 3 is an extreme case of MNCAR, in which missingness is governed by the value of the target variables themselves. This scenario is analogous to test equating situations (Shibayama, 1995), where extreme (minimum or maximum) values tend to be missing on particular variables. MNCAR 2, falling between these two extreme cases, represents a more likely scenario in practical situations.

Table 5 gives a summary of results. Note that the censoring rate is varied from 10% to 30% for the Food-Cancer and Organizational Identification data sets (similarly to the MCAR case), while only from 5% to 15% for Mezzich's data, which makes a direct comparison of this case with the analogous MCAR case rather difficult (except for the 10% censoring case). It can be readily observed that, regardless of the methods, the parameter recovery is affected by the degree of non-randomness in censoring. The recovery generally deteriorates as the degree of non-randomness increases (i.e., correlations between target variables and agent variables increase). As expected, MNCAR 1, deemed closest to MCAR, is least affected, MNCAR 3 is most severely affected, and MNCAR 2 falls between them. The RPCA method seems to work best overall, consistently with the earlier results under the MCAR conditions. This method still suffers from a minor degree of non-randomness in MNCAR 1 relative to the MCAR case, and even more in MNCAR 2 and 3, where non-randomness is more severe. The MDP method is clearly inferior to the RPCA method, which is more pronounced under the MNCAR conditions than under the MNCAR conditions. The DA method is only slightly inferior to the RPCA method in parameter recovery, as in the MCAR cases, but it remains time-consuming to compute. The TSR and WLRA methods fall between the DA methods and

the MDP method, which turns out to be the worst method among all methods tried under the MNCAR conditions.

4 Concluding remarks

In this paper, we compared the performance of five methods for handling missing data in PCA under the MCAR and MNCAR conditions. Specifically, we examined their parameter recovery capability as functions of proportions of missing data, dimensionality of solutions, and the degree of non-randomness in censoring. In the MCAR situations, the results indicated that all methods worked well when the dimensionality and the proportion of missing data were small. Their performance deteriorated as these factors increased, but the speed of deterioration tended to be faster with the WLRA method. The RPCA method has the highest parameter recovery capability regardless of the conditions examined under the present study, and may be regarded as the overall winner. It is iterative, but the computational burden is not too excessive. The DA method comes as a close second in terms of parameter recovery. This method, as implemented here, is extremely time-consuming, however, to the extent that it is unpractical. The TSR method comes as a close third, and it is not too time-consuming to apply. The MDP method is slightly inferior to the top three performers in parameter recovery. All the methods suffered from degrees of non-randomness in censoring processes. Within the MNCAR conditions examined, the RPCA method still worked best, the DA method came as close second, the TSR and WLRA methods close thirds, and the MDP method last.

One general recommendation that may be offered is to keep the number of components as small as dictated by necessity irrespective of the method to be used. Higher dimensional solutions tend to increase the chance of extracting weak components, which invariably works negatively against parameter recovery. It is encouraging to see that some of the methods we tried did reasonably well even under the MNCAR conditions if the censor rate is not excessive (5 to 10%).

The results under the MCAR conditions summarized above are more or less consistent with the previous results (Day and Josse 2015; Folch-Fortuny et al. 2015; van Ginkel et al. 2014), when the conditions (i.e., censor rates, dimensionality of solutions) are comparable. The only exception is the WLRA method, which did not work as well in the present study as in Folch-Fortuny et al.'s study. This may be because Folch-Fortuny et al.'s results were obtained under a fairly lenient stopping criterion in the iterative procedure. Early stopping helps avoid overfitting to missing data, blamed to be the major cause of the poor performance of this method under certain conditions. If so, one may well wonder why not regularly adopt an early stopping rule in the WLRA method. The problem is that there is no good guideline regarding when to stop the iterations in WLRA. The results under the MNCAR conditions, on the other hand, are less comparable to previous studies due to the differences in

Table 5 Recovery of component loadings for missing patterns not completely at random (MNCAR)

			MNCAR 1					MNCAR 2				
Data	dim	p	MDP	WLRA	RPCA	TSR	DA	MDP	WLRA	RPCA	TSR	DA
Food-Cancer	2	10%	1.0	1.0	1.0	1.0	1.0	.95	.98	.99	.99	.94
		20%	.95	1.0	1.0	1.0	.99	.88	.83	.88	.89	.84
		30%	.78	.84	.98	.97	.96	.45	.50	.89	.64	.70
	3	10%	.97	.99	.99	.99	1.0	.93	.90	.99	.97	.89
		20%	.93	.86	.96	.98	.98	.75	.75	.89	.83	.84
		30%	.70	.76	.96	.94	.94	.59	.64	.89	.75	.89
Org. Ident.	3	10%	1.0	.99	1.0	1.0	1.0	.97	.99	.99	1.0	1.0
		20%	.94	.99	.99	.97	.98	.91	.95	.95	.99	.98
		30%	.75	.87	.91	.92	.89	.74	.72	.84	.83	.86
	4	10%	.99	.99	.97	.99	1.0	.95	.98	.98	.99	.99
		20%	.94	.91	.98	.98	.97	.86	.85	.94	.90	.94
		30%	.66	.83	.90	.95	.89	.68	.62	.83	.78	.83
	5	10%	.97	.94	.99	.99	.99	.93	.88	.99	.98	.99
		20%	.87	.81	.95	.94	.96	.83	.85	.89	.94	.91
		30%	.62	.61	.88	.83	.84	.64	.68	.80	.76	.78
Mezzich	3	5%	.99	1.0	1.0	1.0		.94	.99	.99	.99	
		10%	.98	.99	.99	1.0		.82	.95	.93	.97	
		15%	.97	.98	.98	.97		.80	.82	.83	.81	
	4	5%	.97	.99	1.0	1.0		.93	.99	.99	.99	
		10%	.98	.97	.99	.99		.81	.96	.96	.86	
		15%	.94	.92	.98	.96		.77	.74	.81	.79	
	5	5%	.98	.97	.99	1.0		.92	.96	.99	.98	
		10%	.96	.96	.98	.98		.81	.83	.90	.90	
		15%	.85	.86	.97	.95		.75	.74	.76	.74	

			MNCAR 3				
Data	dim	p	MDP	WLRA	RPCA	TSR	DA
Food-Cancer	2	15%	.96	.99	1.0	.99	.99
		20%	.80	.84	.83	.89	.86
		30%	.54	.56	.79	.70	.46
	3	10%	.86	.89	.98	.97	.95
		20%	.67	.73	.78	.70	.80
		30%	.47	.52	.76	.68	.46
Org. Ident.	3	10%	.92	.95	.96	.97	.97
		20%	.85	.87	.91	.94	.92
		30%	.70	.74	.81	.81	.84
	4	10%	.90	.93	.94	.95	.93
		20%	.84	.86	.86	.89	.90
		30%	.64	.66	.76	.78	.84
	5	10%	.80	.90	.91	.92	.92
		20%	.79	.75	.83	.84	.89
		30%	.56	.69	.74	.75	.80
Mezzich	3	5%	.95	.99	.99	.99	
		10%	.80	.84	.88	.87	
		15%	.82	.80	.81	.82	
	4	5%	.94	.99	.99	.99	
		10%	.79	.90	.89	.86	
		15%	.76	.79	.80	.72	
	5	5%	.92	.98	.99	.98	
		10%	.76	.80	.86	.85	
		15%	.73	.77	.81	.76	

generating MNCAR data. Unfortunately, there are no established ways of generating MNCAR data, although in our study we used the correlation between agent and target variables as a measure of non-randomness in missing data patterns. It is reassuring, however, to find that the RPCA method worked best regardless of conditions. The good performance of this method is perhaps due to the fact that the regularization mechanism built in this method provides a good early stopping criterion to avoid overfitting prevalent in the WLRA method.

Rubin (Little and Rubin 1987) further divided conditions of MNCAR into two subcategories, missing at random (MAR) and missing not at random (MNAR), based on whether the agent variables are among the variables subjected to PCA or some outside variables. This distinction does not seem to be so important according to the results obtained. In MNCAR 1, the agent variables are outside variables, but this case is closest to MCAR because they are nearly uncorrelated with the target variables. In MNCAR 2 and MNCAR 3, the agent variables are among those analyzed by PCA, but these cases are less MCAR than MNCAR 1 because the agent variables are more highly correlated with their respective target variables. For parameter recovery, a more crucial factor seems whether enough information is left in observed data after censoring, to construct components which are good approximations to the original components. This observation gets a strong support from studies on variable selection in PCA (e.g., Mori, Iizuka, Tarumi, and Tanaka 2007), which is an "art" of how to deliberately create missing data (discard entire sets of observations on certain variables) in such a way that the original component structures are preserved as much as possible with remaining variables.

Acknowledgements The work reported in this paper has been supported by a research grant (Discovery Grant: 10630) from the Natural Sciences and Engineering Research Council of Canada to the second author. We thank Aida Eslami for providing the reference to Josse and Husson (2012) on RPCA.

References

Bergami M, Bagozzi R P (2000) Self-categorization, affective commitment and group-esteem as distinct aspects of social identity in the organization. Brit J Soc Psychol 39:555–577.

Bernaards C A, Sijtsma K (2000) Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. Multivar Behav Res 35:321–364.

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

Folch-Fortuny A, Arteaga F, Ferrer A (2015) PCA model building with missing data. Chemometr Intell Lab 146:77–88.

Folch-Fortuny A, Arteaga F, Ferrer A (2016) Missing data imputation toolbox for MATLAB. Chemometr Intell Lab 154:93–100.

Gabriel K R, Zamir S (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 22:489–498.

Gifi A (1990) Nonlinear multivariate analysis. Chichester, UK: Wiley.

Grung B, Manne R (1998) Missing values in principal component analysis. Chemometr Intell Lab 42:125–139.
Hwang H, Takane Y (2014) Generalized structured component analysis: A component-based approach to structural equation modeling. Boca Raton, FL: Chapman and Hall/CRC Press.
Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 11:1957–2000.
Josse J, Husson F, Pagès J (2009) Gestion des données manquantes en analyse en composantes principales. Journal de la Société Franccaise de Statistique 150:28–51.
Josse J, Husson F (2012) Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Franccaise de Statistique 153:79–99.
Josse J, Timmerman M E, Kiers H A L (2013) Missing values in multi-level simultaneous component analysis. Chemometr Intell Lab 129:21–32.
Kiers H A L (1997). Weighted least squares fitting using iterative ordinary least squares algorithms. Psychometrika 62:251–266.
Little R J A, Rubin D B (1987) Statistical analysis with missing data. New York: Wiley.
McDonald R P, Burr E J (1967) A comparison of four methods of constructing factor scores. Psychometrika 32:381–401.
Meulman J J (1982) Homogeneity analysis of incomplete data. Leiden, The Netherlands: DSWO Press.
Mezzich J E (1978) Evaluating clustering methods for psychiatric diagnosis. Biol Psychol 13:265–281.
Mori Y, Iizuka M, Tarumi T, Tanaka Y (2007) Variable selection in principal component analysis. In W Härdle, Y Mori, P Vieu (eds.) Statistical mehtods for biostatistics and related fields (pp. 265–283). Berlin: Springer.
Overall J E, Gorham D R (1962) The brief psychatric rating scale. Psychol Rep 10:799–812.
Rubin D B (1987) Multiple imputation for nonresponse in survey. New York: Wiley.
Schafer J L (1997) Analysis of incomplete multivariate data. New York: Wiley.
Segi M (1979) Age-adjusted death rates for cancer for selected sites (A-classification) in 51 countries in 1974. Nagoya, Japan: Segi Institute of Cancer Epidemiology.
Serneels S, Verdonck T (2008) Principal component analysis for data containing outliers and missing elements. Computational Statistics and Data Analysis 52:1712–1727.
Shibayama T (1995) A linear composite method for test scores with missing values. Memoirs of the Faulty of Eduction, Niigata Univerity 36:445–455.
Stanimirova I, Daszykowski M, Walczak B (2008) Dealing with missing values and outliers in principal component analysis. Talanta 72:172–178.
Takane Y (2013) Constrained principal component anlysis and related techniques. Boca Raton, FL: Chapman and Hall/CRC Press.
Takane Y, Oshima-Takane Y (2003) Relationships between two methods for dealing with missing data in principal component analysis. Behaviometrika 30:145–154.
Tanner M A, Wong W H (1987) The calculation of posterier distributions by data augumentation (with discussion). J Am Stat Assoc 82:528–550.
Tipping M E, Bishop C M (1999) Probabilistic principal component analysis, J Roy Stat Soc B 61:611–622.
Tucker L R (1951) A method of synthesis of factor analysis studies. Personnel Research Section Report No. 984, U. S. Department of Army, Wasgington, D. C.
Van Ginkel J R, Kroonenberg P M (2014) Using generalized procrustes analysis for multiple imputation in principal component analysis. J Classif 31:242–269.
Van Ginkel J R, Kroonenberg P M, Kiers H A L (2014) Missing data in principal component analysis of questionnaire data. J Stat Comput Sim 84: 2298–2315.
Walczak B, Massart D L (2001) Dealing with missing data, Part 1. Chemometr Intell Lab 58:15–27.
Wentzell P D, Andrews D T, Hamilton D C, Faber K, Kowalski B R (1997) Maximum likelihood principal component analysis. J Chemomet 11:339–366.

The following two tables accompany the paper as material in Online Resource.

Table A.1 Componentwise recovery of loadings for the organizational identification data: Mean congruence coefficients as functions of the censor rate and the method

	No. of dimensions	3			4			
p	Method\Comp.	1	2	3	1	2	3	4
	MDP	1.0	1.0	.98	1.0	1.0	.98	.95
	WLRA	1.0	1.0	.99	1.0	1.0	.98	.96
10%	RPCA	1.0	1.0	.99	1.0	.99	.99	.97
	TSR	1.0	1.0	.99	1.0	1.0	.99	.97
	DA	1.0	1.0	.99	1.0	1.0	.99	.97
	MDP	1.0	.99	.95	1.0	1.0	.95	.87
	WLRA	1.0	.99	.96	.97	.95	.91	.80
20%	RPCA	1.0	1.0	.98	1.0	1.0	.98	.90
	TSR	1.0	1.0	.97	1.0	1.0	.97	.93
	DA	1.0	1.0	.98	1.0	1.0	.98	.91
	MDP	1.0	.99	.90	1.0	.99	.91	.74
	WLRA	1.0	.99	.93	.94	.89	.81	.71
30%	RPCA	1.0	1.0	.95	1.0	.99	.95	.85
	TSR	1.0	1.0	.95	1.0	1.0	.95	.85
	DA	1.0	1.0	.95	1.0	1.0	.95	.79

	No. of dimensions	5				
p	Method\Comp.	1	2	3	4	5
	MDP	1.0	1.0	.98	.94	.92
	WLRA	.98	.96	.93	.87	.86
10%	RPCA	1.0	1.0	.99	.97	.96
	TSR	1.0	1.0	.99	.97	.95
	DA	1.0	1.0	.99	.97	.95
	MDP	1.0	.99	.95	.87	.79
	WLRA	.94	.86	.75	.72	.66
20%	RPCA	1.0	1.0	.98	.91	.88
	TSR	1.0	1.0	.97	.92	.88
	DA	1.0	1.0	.97	.92	.88
	MDP	1.0	.99	.90	.79	.67
	WLRA	.74	.64	.63	.53	.50
30%	RPCA	1.0	.99	.95	.85	.76
	TSR	1.0	.99	.95	.87	.78
	DA	1.0	.99	.95	.85	.76

	No. of dimensions	6					
p	Method\Comp.	1	2	3	4	5	6
	MDP	1.0	1.0	.98	.94	.92	.77
	WLRA	.95	.91	.84	.78	.73	.68
10%	RPCA	1.0	1.0	.99	.97	.96	.86
	TSR	1.0	1.0	,99	.97	.95	.89
	DA	1.0	1.0	.99	.97	.95	.87
	MDP	1.0	.99	.95	.88	.81	.59
	WLRA	.80	.72	.64	.58	.52	.47
20%	RPCA	1.0	1.0	.98	.91	.89	.73
	TSR	1.0	1.0	.97	.91	.89	.71
	DA	1.0	1.0	.99	.92	.89	.75
	MDP	1.0	.99	.89	.80	.49	.52
	WLRA	.53	.46	.49	.39	.31	.32
30%	RPCA	1.0	.99	.94	.85	.78	.58
	TSR	1.0	.99	.94	.86	.79	.60
	DA	1.0	1.0	.95	.85	.79	.59

Table A.2 Componentwise recovery of loadings for Mezzich's data: Mean congruence coefficients as functions of the censor rate and the method

	No. of dimensions	2		3		
p	Method\Comp.	1	2	1	2	3
	MDP	1.0	.99	1.0	.99	.98
10%	WLRA	1.0	.99	1.0	.99	.99
	RPCA	1.0	.99	1.0	.99	.97
	TSR	1.0	1.0	1.0	.98	.95
	MDP	.99	.96	.99	.97	.96
20%	WLRA	1.0	.98	1.0	.98	.97
	RPCA	.99	.97	.99	.97	.93
	TSR	.99	.96	.99	.96	.93
	MDP	.99	.94	.99	.94	.91
30%	WLRA	.99	.96	.99	.96	.94
	RPCA	.98	.92	.98	.93	.85
	TSR	.99	.91	.97	.89	.86

	No. of dimensions	4			
p	Method\Comp.	1	2	3	4
	MDP	1.0	.99	.98	.90
10%	WLRA	1.0	.99	.99	.95
	RPCA	1.0	.99	.98	.97
	TSR	1.0	.99	.97	.96
	MDP	.99	.97	.95	.75
20%	WLRA	.97	.95	.94	.84
	RPCA	.99	.97	.94	.92
	TSR	.99	.96	.92	.90
	MDP	.99	.95	.92	.66
30%	WLRA	.90	.82	.80	.73
	RPCA	.93	.88	.83	.81
	TSR	.95	.86	.80	.74

	No. of dimensions	5				
p	Method\Comp.	1	2	3	4	5
	MDP	1.0	.99	.98	.91	.81
10%	WLRA	.99	.98	.97	.92	.88
	RPCA	1.0	.99	.98	.97	.93
	TSR	1.0	.98	.97	.96	.92
	MDP	.99	.97	.95	.83	.61
20%	WLRA	.94	.84	.82	.79	.69
	RPCA	.99	.97	.94	.92	.80
	TSR	.98	.93	.87	.85	.71
	MDP	.99	.94	.92	.90	.53
30%	WLRA	.68	.65	.55	.55	.54
	RPCA	.98	.93	.87	.82	.58
	TSR	.98	.78	.70	.61	.47

	No. of dimensions	6					
p	Method\Comp.	1	2	3	4	5	6
	MDP	1.0	.99	.98	.92	.84	.75
10%	WLRA	.99	.96	.95	.90	.83	.80
	RPCA	1.0	.99	.97	.97	.93	.87
	TSR	.99	.97	.97	.95	.89	.79
	MDP	.99	.97	.96	.81	.71	.58
20%	WLRA	.85	.76	.73	.67	.62	.60
	RPCA	.99	.97	.93	.91	.81	.63
	TSR	.96	.87	.80	.77	.64	.47
	MDP	.99	.94	.90	.72	.62	.48
30%	WLRA	.48	.47	.41	.30	.32	.30
	RPCA	.98	.91	.84	.78	.60	.46
	TSR	.83	.65	.56	.52	.37	.31