

Chapter 8

Handling Missing Data in Principal Component Analysis Using Multiple Imputation



Joost R. van Ginkel

Abstract Principal component analysis (PCA) is a widely used tool for establishing the dimensional structure in questionnaire data. Whenever questionnaire data are incomplete, the missing data need to be treated prior to carrying out a PCA. Several methods exist for handling missing data prior to carrying out a PCA. The current chapter first discusses the most recent developments regarding the treatment of missing data in PCA. Next, of these methods, the method that is most promising both from a theoretical and practical point of view will be discussed in more detail, namely, multiple imputation. Finally, some extensions of multiple imputation to other PCA-related techniques or to statistics within PCA beyond the basics are discussed, and some general recommendations regarding the use of PCA on multiply imputed datasets in different statistical software packages will be given.

8.1 Introduction

One important part of establishing the psychometric properties of a test or questionnaire is determining its dimensional structure. Oftentimes measurement instruments measure different aspects of the same psychological construct. For example, a questionnaire may measure different ways in which one can be religious (Hills et al., 2005) or different aspects of schizotypal personality disorder (Mata et al., 2005).

Although establishing the dimensional structure of a measurement instrument is mostly done in personality assessment, there are also situations in educational settings where dimensionality of a measurement instrument may be relevant. For example, in a school setting, one may be interested in students' attitudes towards different types of bullying (Boulton et al., 1999) or different aspects of students' well-being (Borgonovi & Pál, 2016). As the developer of such measurement instruments, you may want to know whether its items indeed measure the specific aspect

J. R. van Ginkel (✉)
Methodology and Statistics, Leiden University, Leiden, The Netherlands
e-mail: jginkel@fsw.leidenuniv.nl

of the trait that they are intended to measure. In such cases a statistical technique is used that establishes which items measure which aspect of the underlying construct. One widely used technique for this purpose is *principal component analysis* (PCA).

In practice, many datasets that are used for determining the dimensional structure of questionnaires suffer from missing data. When data are incomplete, this complicates the use of PCA or any analyses that are aimed towards determining the dimensional structure. When missing data are not properly handled, erroneous conclusions may be drawn about dimensional structure of the measurement instrument. It is therefore important that missing data are properly treated prior to determining the dimensional structure.

The current chapter is going to focus on a situation where one is interested in determining the dimensional structure of a test or questionnaire using PCA in an incomplete dataset. In the first part of this chapter, an overview of the most recent developments of missing data handling in PCA will be given. In this overview, several methods for handling missing data in PCA are going to be discussed. The second part will focus on the method that is the most promising one both from a theoretical and practical point of view in more detail: multiple imputation. The chapter will end with some extensions of missing data handling in PCA to statistics within PCA beyond the basics and to PCA-related techniques, and some general recommendations regarding the use of PCA on multiply imputed datasets in several statistical software packages are given.

8.2 Principal Component Analysis

Within a questionnaire, different subsets of items may exist that each are supposed to measure a different aspect of the same construct. Such a subset is also called a subscale. In PCA, the goal is to reduce a large number of continuous variables J to a smaller number of components, K . Although theoretically the variables need to be continuous, in practice PCA is regularly applied to items measured on a Likert scale.

Suppose \mathbf{Z} is the standardized dataset consisting of the responses of I respondents to J items. In PCA, by means of a singular value decomposition, \mathbf{Z} is decomposed as:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \quad (8.1)$$

Here, \mathbf{U} is a column wise orthonormal $N \times J$ matrix, \mathbf{V} is a column wise orthonormal $J \times J$ matrix, and $\mathbf{\Lambda}$ is a $J \times J$ diagonal matrix with the singular values on the main diagonal. The singular values are the square roots of the eigenvalues. An important part of the output in PCA that gives insight in how the items in the data are related to the different underlying components is the $J \times J$ component matrix. This matrix is computed as $\mathbf{A} = N^{-1/2}\mathbf{V}\mathbf{\Lambda}$ and contains the correlations between the variables and the components. These correlations coincide with the regression

coefficients (loadings) from multivariate multiple regression of the item scores on the principal components.

In the original singular value decomposition, there are as many components as there are variables. However, usually only the first few components explain a substantial portion of the variance of the variables in \mathbf{Z} . Additionally, given that a goal in PCA is to reduce the original number of variables J to a smaller set of dimensions K ($K < J$) and given that in PCA the dimensions are represented by the components, usually only a smaller number of components K are used for interpretation (there are several ways for determining K . See, for example, Furr, 2018, pp. 85–92). The resulting reduced component matrix is denoted by \mathbf{A}_K ($J \times K$).

For interpretational purposes the resulting \mathbf{A}_K matrix may be rotated using either Varimax rotation or Oblique rotation (Harman, 1976). The rotated component matrix is denoted \mathbf{A}_K^* .

8.3 Missing Data

As already mentioned in the introductory section, in the data collection process, it may happen that not all respondents provide answers to all the questions in the questionnaire. Reasons for this may be that a respondent finds a question too personal, that (s)he accidentally skipped a question, (s)he did not understand the question, and so on. When respondents have not answered all the questions, this results in a dataset with missing data.

When data are incomplete, this might have consequences for the PCA that is carried out next. Before the PCA can be carried out, the missing data need to be handled. Several ways to deal with missing data in PCA exist (to be discussed later on), ranging from very simple to highly advanced. However, each of these methods makes either explicit or implicit assumptions about the underlying process that caused the missing data, also called the *missingness mechanism*. Rubin (1976) and Little and Rubin (2002) defined three main missingness mechanisms, namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). As these missingness mechanisms are extensively described by Rubin (1987), Little and Rubin (2002), and various other literature on missing data, they will only briefly be discussed here.

8.3.1 Missingness Mechanisms

When the data are MCAR, there is no relation between the missing values and any observed or unobserved information. Consequently, the missing data are randomly scattered across the dataset. Under MAR, missing data may depend on observed data but not on unobserved data. It could be, for example, that within different age

groups, respondents have different amounts of missing data on the questions in the questionnaire. If, however, age is observed for all respondents and within each age group the missing data are randomly scattered across the data, then the missingness is MAR. Finally, MNAR is any missingness mechanism that does not qualify as either MCAR or MAR. Thus, under MNAR the missingness depends either on a variable that was not included in the data collection process (e.g., the older people get, the more missing data they have, and age not being observed) or on the value of the missing score itself (people with higher scores on an item in the questionnaire being more likely not to answer the item than people with low scores), or both.

8.3.2 *Methods for Handling Missing Data*

Methods for dealing with missing data in PCA range from ad hoc to highly advanced. Among the ad hoc procedures are *listwise deletion* and *pairwise deletion*; a few examples of advanced methods are *missing data passive* (Meulman, 1982; Takane & Oshima-Takane, 2003), *regularized PCA* (Josse et al., 2009), *EM-covariances* (Bernaards & Sijtsma, 2000), and *multiple imputation* (Rubin, 1987; Van Ginkel & Kroonenberg, 2014). Although the methods mentioned here are not exhaustive, with the exception of listwise deletion and pairwise deletion, they all have in common that they are advanced in the sense that they all carry out the PCA in a statistically sound way, without throwing away any data.

In the abovementioned references, usually the performance of only one of these methods was compared with the performance of other less advanced methods (such as substituting the variable mean for each missing value) or with different variants of the same method. However, none of these studies compared all of these advanced methods with each other. Van Ginkel et al. (2014) did a simulation study in which they compared all of the abovementioned methods. Before discussing the results of their study, each of these methods will be discussed in more detail first. In so doing they will be categorized into three categories, namely, *traditional methods*, *simultaneous methods*, and *sequential methods*.

8.3.2.1 **Traditional Methods**

The traditional methods described by Van Ginkel et al. (2014) are listwise deletion and pairwise deletion. Listwise deletion deletes every case with at least one missing value on any of the variables in the PCA from the analysis. Since usually more data points are thrown away than there are missing data points, listwise deletion is very wasteful. An additional problem of listwise deletion is that in general, unbiased results of statistical analyses are only guaranteed when the data are MCAR. However, in PCA, component loadings are intrinsically biased. This has to do with the fact that they are bound to -1 and $+1$, as in normal correlations (Fisher, 1915). Consequently, in PCA the question is not whether loadings are biased as a

result of listwise deletion, but how much more biased they are than without missing data.

Like listwise deletion, pairwise deletion deletes cases with missing data for the calculation of the component loadings, but in doing so it uses more observed information than listwise deletion does. Pairwise deletion calculates the component loadings from a PCA in a slightly different way than is described in Sect. 8.2. Rather than carrying out a singular value decomposition on the standardized dataset, pairwise deletion computes the component loadings by performing an eigenvalue decomposition on the correlation matrix (technical details are discussed in Tabachnick & Fidell, 2001, pp. 591–595). In doing so it also deletes cases with missing values, but does this for each variable pair for which a correlation is computed, separately. Consequently, in pairwise deletion more information is used than in listwise deletion.

Although pairwise deletion uses more information from the data than listwise deletion does, an implicit assumption is still that the data are MCAR. An additional disadvantage of pairwise deletion is that since each correlation is based on different cases, combinations of correlations may occur that together form a correlation matrix that is not positive semi-definite. Consequently, computational problems may occur when computing the component loadings.

8.3.2.2 Simultaneous Methods

Van Ginkel et al. (2014) discussed two methods that estimate the loadings of the PCA and handle the missing data in the process, namely, *missing data passive* (Meulman, 1982; Takane & Oshima-Takane, 2003) and *regularized PCA* (Josse et al., 2009). Since both methods estimate the PCA and in the process also handle the missing data while not throwing away any information, these methods were referred to as *simultaneous methods*.

The idea of missing data passive is that a weight matrix of 1's (observed data) and 0's (unobserved data) is used in a weighted homogeneity analysis, a categorical form of PCA. Regularized PCA, on the other hand, is based on PCA using weighted least squares (Kiers, 1997; Grung & Manne, 1998). In weighted least squares, after filling starting values for the missing data, an iterative algorithm is used that alternates between a regression analysis predicting the component scores from the current estimates of the loadings and a regression analysis predicting the loadings from the current estimates of the component scores. At each iteration, the estimates for the missing data are updated. Regularized PCA is based on the same principle. The difference with weighted least squares is that regularized PCA uses a smoothing procedure for estimating the missing data in the process. This smoothing procedure is especially useful when many components are extracted as weighted least squares may break down in case of many components.

The simultaneous methods have two theoretical advantages over pairwise deletion. Firstly, they do not throw away data like pairwise deletion does. Secondly, as long as the missing data are related to variables that take part in the PCA, using

these methods will not introduce any additional bias in the component loadings as a result of deviations from MCAR.

8.3.2.3 Sequential Methods

Lastly, Van Ginkel et al. (2014) discussed two methods that treat the missing data separately from the calculation of the component loadings: *EM-covariances* (Bernaards & Sijtsma, 2000) and *multiple imputation* (Rubin, 1987). In EM-covariances, first an expectation-maximization algorithm (EM; Dempster et al., 1977) is used to obtain full information maximum likelihood estimates of the means and covariances of the data under the assumption that the data are multivariate normally distributed. Next, the covariances of the variables that are part of the PCA are converted to correlations, and an eigenvalue decomposition of this correlation matrix is carried out to obtain the component loadings.

EM-covariances has the same theoretical advantages over pairwise deletion that missing data passive and regularized PCA have. However, whereas missing data passive and regularized PCA can only handle MAR mechanisms where the missing data depend on variables that are included in the PCA, EM-covariances can also handle MAR mechanisms where the missingness depends on variables outside the PCA, as long as they are included in the maximum likelihood estimation of the covariance matrix.

Multiple imputation is perhaps the most widely recommended method for dealing with missing data. This procedure works in three steps. In the first step, the missing data are estimated multiple (M) times according to a statistical model that accurately describes the structures present in the data. This results in M complete versions of the incomplete dataset, which only differ in the estimates for the missing data. In the second step, the statistical analysis of interest is applied to each of the M completed datasets, resulting in M different outcomes of the same analysis (in the current context, a PCA). Finally, the results of the M analyses are combined into one overall result, using specific calculations, denoted *combination rules* (for the specific PCA context, combination rules will be discussed in Sects. 8.5.1 and 8.5.2).

Like EM-covariances, multiple imputation can handle any MAR mechanism, regardless of whether the missingness depends on a variable within the PCA or outside the PCA. However, an additional advantage of multiple imputation is that the multiply imputed data can be used for almost any type of statistical analysis other than PCA, whereas the means and covariance matrices of EM-covariances can only be used as the input for analyses that use means and covariances.

8.3.2.4 Which Method for Handling Missing Data in PCA Is the Preferred One?

In this subsection a short summary of the results found by Van Ginkel et al. (2014) is given. Based on the results and on the theoretical properties of each method, a recommendation is given on which method is generally the best one to use.

To determine the performance of each method, Van Ginkel et al. (2014) studied three quality measures in their simulation study, namely, the root mean squared bias (RMSB) of the component loadings, the mean bias (MB) of the component loadings, and the average number of items assigned to the incorrect component, denoted the *classification error* (CE). The RMSB, MB, and CE were defined as follows: Suppose that a_{jk}^* is the population component loading of item j on Varimax rotated component k and $\hat{a}_{jk,d}^*$ is the corresponding loading for the incomplete simulated dataset d ($d = 1, \dots, D$) in a specific condition of the simulation study (specific missing data handling method, specific percentage of missingness, etc.). For the specific condition, the RMSB is:

$$\text{RMSB} = \frac{1}{D} \sum_{d=1}^D \sqrt{\frac{\sum_{j=1}^J \sum_{k=1}^K (\hat{a}_{jk,d}^* - a_{jk}^*)^2}{JK}}, \tag{8.2}$$

and the MB is:

$$\text{MB} = \frac{1}{JKD} \sum_{d=1}^D \sum_{j=1}^J \sum_{k=1}^K (\hat{a}_{jk,d}^* - a_{jk}^*). \tag{8.3}$$

As for the CE, define f as the component number of the component for which it holds that

$$a_{jf}^* = \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right)$$

and g as the component number of the component for which it holds that

$$\hat{a}_{jg,d}^* = \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right).$$

Next, based on guidelines by Comrey and Lee (1992) that state that loadings below 0.32 should not be interpreted, define:

$$\begin{aligned} w_{j,d} &= 0 \text{ if } \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right) < 0.32 \text{ and } \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right) < 0.32 \\ w_{j,d} &= 0 \text{ if } \max \left(|a_{j1}^*|, \dots, |a_{jK}^*| \right) > 0.32 \text{ and } \max \left(|\hat{a}_{j1,d}^*|, \dots, |\hat{a}_{jK,d}^*| \right) > 0.32 \\ &\text{and } f = g \\ w_{j,d} &= 1 \text{ otherwise.} \end{aligned}$$

For the specific condition, the CE is:

$$CE = \frac{1}{D} \sum_{d=1}^D \sum_{j=1}^J w_{j,d}. \quad (8.4)$$

The results of the traditional methods will be discussed first. Van Ginkel et al. (2014) studied the performance of all methods under both MCAR, MAR, and MNAR. The bias of the individual component loadings was not studied so it remains unclear how much bias in component loading deviations from MAR introduced for the advanced methods and how much bias deviations from MCAR introduced for listwise deletion and pairwise deletion. However, it became clear from the study that listwise deletion did not perform well on either the *RMSB*, the *MB*, or the *CE*, regardless of the missingness mechanism. Additionally, for high percentages of missing data, listwise deletion was not even feasible because after removing the incomplete cases, no or too few complete cases were left to analyze. In short, based on the results of Van Ginkel et al. (2014), listwise deletion is not recommended for PCA.

As for pairwise deletion, Van Ginkel et al. (2014) found that this method actually gave satisfactory results on all three quality measures, regardless of the missingness mechanism. Additionally, computational problems did not occur in the situations studied by Van Ginkel et al. (2014). However, the latter does not mean that these problems cannot occur in practice, so using pairwise deletion in practice may not always be feasible.

Regarding the simultaneous methods, Van Ginkel et al. (2014) found that, firstly, missing data passive generally gave results that were similar to pairwise deletion with respect to the outcome measures and that missingness mechanism did not have a substantial effect on the performance of missing data passive. Regularized PCA, on the other hand, produced results that were slightly worse than those of pairwise deletion and missing data passive. Thus, despite their theoretical advantages over pairwise deletion, they do not seem to show in the quality measures in the study by Van Ginkel et al. (2014).

Finally, regarding the sequential methods, Van Ginkel et al. (2014) found that regarding the outcome measures multiple imputation and EM-covariances performed similar to pairwise deletion. Thus, despite the theoretical advantages of multiple imputation and EM-covariances over the other methods, this does not really seem to show in the quality measures either. This leaves us with the question which method is the preferred one.

Of all the methods discussed in the previous subsections, multiple imputation is the method that is most preferred *from a theoretical point of view* because it will not introduce additional bias in component loadings under any MAR mechanism. Not considering lower benchmark listwise deletion, pairwise deletion is the least preferred method from a theoretical point of view because it assumes MCAR, and it may run into computational problems. However, Van Ginkel et al. (2014) showed that although multiple imputation was one of the better performing methods, it did not perform any better than pairwise deletion. Furthermore, pairwise deletion

(together with listwise deletion) is the simplest method for handling missing data in PCA (it is included in most statistical software packages and does not require any additional preprocessing of the data). This raises the question whether pairwise deletion should not be preferred over any missing data handling method for PCA at all times, including multiple imputation. Not quite, as being simple and performing well on a number of outcome measures may not necessarily be the only criteria for preferring a missing data method in PCA over others. There are other things that may have to be taken into consideration as well.

Firstly, even though it did not occur in the simulation study by Van Ginkel et al. (2014), in practice computational problems may still occur using pairwise deletion. Secondly, even when computational problems do not occur, then still the question is what the sample size is that the PCA solution is based on as some correlations are computed for different cases and different number of cases than others.

Thirdly, sometimes researchers may be interested in confidence intervals of principal component loadings. Van Ginkel and Kiers (2011) developed ways to construct bootstrap confidence intervals for component loadings in multiply imputed datasets (more on this in Sect. 8.4) and showed that these ways performed well regarding coverage of the population loadings. For pairwise deletion there is no way to construct bootstrap confidence intervals of the component loadings.

Finally, and most importantly, a practical advantage of multiple imputation over all other methods for handling missing data including pairwise deletion is that multiple imputation provides the researcher a complete dataset which can be used for other statistical analyses as well. In practice, a dataset is almost never subjected to one single statistical analysis, so it is desirable to have a general solution for all analyses that are carried out on the dataset, such that all analyses on these datasets are comparable regarding sample size, regarding the cases used, and regarding the data points (both observed and imputed). When not imputing the data and analyzing only the usable data however, for some analyses listwise deletion will be applied (and for each of these analyses, different cases may be used, depending on which variables are included in the specific analysis), for other analyses full information maximum likelihood will be applied, and yet for other analyses, pairwise deletion (as in PCA) will be applied. This will make the statistical analyses mutually incomparable.

Additionally, while pairwise deletion may give good results for PCA, this is not necessarily the case for other analyses that are applied to the dataset. It has been well established that multiple imputation performs better regarding bias and coverage of parameters than methods based on deleting data (listwise/pairwise deletion). Consequently, when a researcher decides not to impute the data, conclusions regarding PCA may be valid, but conclusions based on other statistical analyses on the same dataset may not.

In short, although from the study of Van Ginkel et al. (2014) we cannot conclude that multiple imputation necessarily recovers the PCA solution better than pairwise deletion does, there are numerous other advantages of multiple imputation over pairwise deletion in PCA. For the remainder of this chapter we are hence going to take the standpoint that multiple imputation is to be recommended most for handling

missing data in PCA. Hence, we are going to get into more detail about multiple imputation in the context of PCA in the next section.

8.4 Multiple Imputation in Principal Component Analysis

As already said in Sect. 8.3.2.3, multiple imputation works in three steps: (1) the imputation step, where multiple estimates for the missing data are generated; (2) the analysis step, where each of the resulting M imputed datasets is analyzed using the statistical analysis of interest; and (3) the combination of the M results into one overall result. Various methods for generating multiple estimates of the missing data in step 1 have been developed, and various texts have been written on them (e.g., Schafer, 1997; Van Buuren, 2018). The general process of generating multiple imputed values for the missing data is not tied to PCA as an analysis for the data, but is generally the same for all statistical analyses that follow after the data have been multiply imputed. Consequently, technical details regarding the process of generating multiple imputed values are not further discussed here. The interested reader is referred to Van Buuren (2018).

In the context of PCA, the second step in the multiple imputation process is carrying out a PCA on each of the M complete versions of the incomplete dataset. This step has already been explained in Sect. 8.2 so this step will not be discussed here either. This leaves us with the third and final step of the multiple imputation process: the combination of M PCA results into one overall PCA result. Van Ginkel and Kroonenberg (2014) discussed combination techniques for the results of PCA in multiply imputed data, which will be discussed next.

8.4.1 Combining the Component Loadings

8.4.1.1 The Problem of Traditional Combination Rules When Applied to PCA

Once a PCA has been obtained from each of the M imputed datasets, this leaves us with M sets of component loadings. The question is how these component loadings are combined into one overall set of component loadings. Rubin (1987) defined combination rules for a parameter estimate with its statistical test and confidence interval. An overall parameter estimate is obtained by averaging the M estimates of the parameter. Considering a component loading $a_{jk,m}$ on variable j on component k to be a parameter estimate of imputed dataset m , a direct application of Rubin's combination rules for parameter estimates would come down to averaging the M component loadings $a_{jk,m}$.

Van Ginkel and Kroonenberg (2014) argued that averaging component loadings across M imputed datasets has three potential problems. Firstly, the order of the

components may not be the same for all M imputed datasets. For example, in one imputed dataset, a set of items may load highest on the first component, while in another imputed dataset, this same set may load highest on the second component. This may especially happen when two adjacent components have near equal variance.

Secondly, many questionnaires contain both indicative items (a higher score means a higher amount of the underlying construct) and contraindicative items (a higher score means a lower amount of the underlying construct). When a specific subscale of a questionnaire contains about as many indicative items as contraindicative items, it could happen that in one or more imputed datasets, the signs of the loadings are reversed compared to those of the other imputed datasets. When averaging these loadings, their signs may cancel each other out, resulting in an average loading lower than the average of the absolute values.

A third disadvantage is that even when sign changes of loadings switching of the order of components do not occur among the M $\mathbf{A}_{K,m}$ matrices, then still the M matrices are not optimally aligned as a result of rotational freedom. Because of this rotational freedom, the average solution is computed across solutions that have more variation among each other than necessary (e.g., Chatterjee, 1984; Markus, 1994; Milan & Whittaker 1995; Linting et al. 2007).

8.4.1.2 Using Generalized Procrustes Analysis for Combining the Component Loadings

A procedure that can resolve all of the three abovementioned problems is Generalized Procrustes analysis (Ten Berge, 1977; Gower, 1975). Generalized Procrustes analysis was originally proposed to derive one overall component solution from several ones, not necessarily obtained from multiply imputed data (e.g., from several different studies). However, Van Ginkel and Kroonenberg (2014) proposed this procedure to explicitly combine the results of several PCA solutions obtained from M imputed datasets. In a simulation study, they showed that this method gave better results regarding *RMSB* (see Eq. (8.2)) than averaging of component loadings did.

In the context of M PCA solutions obtained from M imputed datasets, generalized Procrustes analysis works as follows. Suppose that we have unrotated component matrix $\mathbf{A}_{K,m}$ of imputed dataset m ($m = 1, \dots, M$). We need an orthogonal $K \times K$ rotation matrix \mathbf{T}_m for each of the M imputed datasets that minimizes the sum of squared distances between the transformed loading matrices, given by:

$$f(\mathbf{T}_1, \dots, \mathbf{T}_M) = \sum_{i < j} \text{tr}(\mathbf{A}_{K,i}\mathbf{T}_i - \mathbf{A}_{K,i}\mathbf{T}_i)' (\mathbf{A}_{K,j}\mathbf{T}_j - \mathbf{A}_{K,j}\mathbf{T}_j). \quad (8.5)$$

The rotation matrices $\mathbf{T}_1, \dots, \mathbf{T}_M$ are obtained using a procedure that is a generalization of the classical orthogonal Procrustes problem (Green, 1952; Gower, 1971). In the classical Procrustes problem, we have two matrices \mathbf{A} and \mathbf{B} where \mathbf{A}

needs to be optimally rotated to \mathbf{B} . The required rotation matrix for this problem is found as follows: suppose \mathbf{QLV}' is the singular value decomposition of matrix $\mathbf{A}'\mathbf{B}$. The rotation matrix \mathbf{T} is obtained by:

$$\mathbf{T} = \mathbf{QV}' \quad (8.6)$$

Finally, \mathbf{A} can be optimally rotated to \mathbf{B} by post-multiplying \mathbf{A} by \mathbf{T} .

When optimally rotating M component matrices towards each other, we can use an algorithm by Ten Berge (1977, p. 272). Suppose t is the iteration number, and starting at $t = 1$, the algorithm has the following steps:

Step 0: Set $\mathbf{T}_m = \mathbf{I}$ for $m = 2, \dots, M$.

Step 1: Rotate $\mathbf{A}_{K,1}$ optimally to $\mathbf{B} = \sum_{m=2}^M \mathbf{A}_{K,m} \mathbf{T}_m$ using rotation matrix \mathbf{T}_1 as computed in the right-hand side of Eq. (8.6), yielding $\mathbf{A}_{K,1} \mathbf{T}_1^{(t)}$.

Step 2: Rotate $\mathbf{A}_{K,2}$ optimally to $\mathbf{B} = \mathbf{A}_{K,1} \mathbf{T}_1^{(t)} + \sum_{m=3}^M \mathbf{A}_{K,m} \mathbf{T}_m$, yielding $\mathbf{A}_{K,2} \mathbf{T}_2^{(t)}$.

Step M : Rotate $\mathbf{A}_{K,M}$ optimally to $\sum_{m=1}^{M-1} \mathbf{A}_{K,m} \mathbf{T}_m^{(t)}$, yielding $\mathbf{A}_{K,M} \mathbf{T}_M^{(t)}$.

Step $M + 1$: Rotate $\mathbf{A}_{K,1} \mathbf{T}_1^{(t)}$ optimally to $\mathbf{B} = \sum_{m=2}^M \mathbf{A}_{K,m} \mathbf{T}_m^{(t)}$, yielding $\mathbf{A}_{K,1} \mathbf{T}_1^{(t+1)}$.

Next, the steps 2– M are repeated, where t increases with 1 at each iteration, until convergence. Once convergence has been achieved, the mean of all transformed solutions, also denoted the *centroid* solution $\mathbf{A}_{K,C}$, is used as the pooled PCA solution for the M imputed datasets. Like a PCA solution in complete data, $\mathbf{A}_{K,C}$ can be rotated either with an orthogonal or an oblique transformation.

8.4.2 Uncertainty About the Component Loadings

In the traditional way in which PCA is used, usually no statistical tests or confidence intervals are computed. There are procedures for confidence intervals of population component loadings (more on this in Sect. 8.5), but normally PCA is mainly used without any statistical testing.

However, in multiple imputation uncertainty is created about parameter estimates by the fact that for each imputed dataset the imputed values differ and that this results in slightly different sets of PCA loadings for each imputed dataset. Although $\mathbf{A}_{K,C}$ gives an impression of what the actual sample loadings without missing data would have been, there is still uncertainty about this centroid solution as a result of the variation of the imputed values.

Van Ginkel and Kroonenberg (2014) discussed a procedure to show variation in the component loadings as a result of imputation uncertainty. Using this procedure a loading plot of one component against the other is created, which shows both the centroid solution represented by dots and the uncertainty of the centroid solution

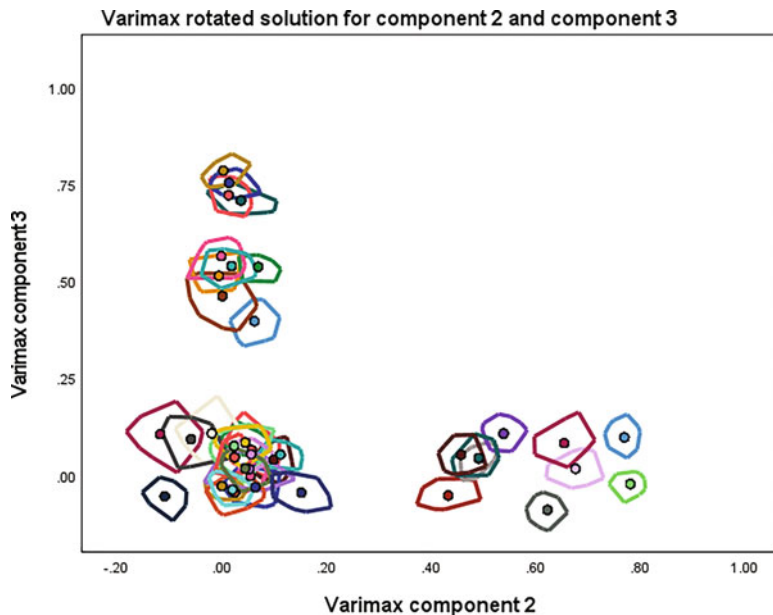


Fig. 8.1 Loading plot of a Varimax rotated four-component solution of components 2 and 3, applied to a multiply imputed dataset with $M = 100$ imputations. The loading plot shows both the centroids and their convex hulls

represented by areas surrounding the dots. These areas are called convex hulls. Figure 8.1 displays a loading plot that includes both the centroid solution of the M PCAs and the convex hulls.

The surface of the convex hulls may serve as a measure of uncertainty about the PCA loadings. These surfaces may be computed in the following way. Each convex hull may be decomposed as several triangles. Suppose a triangle has three sides, namely, a , b , and c , and we define $s = (a + b + c)/2$. See Fig. 8.2. By using Heron's rule dating back to before 200 BC, the surface of one triangle can be determined as $\sqrt{[s(s-a)(s-b)(s-c)]}$. Doing this for all triangles that the convex hull is composed of, and adding up the surfaces, the total surface of the convex hull is obtained.

It should be noted that the convex hulls do not in any way intend to represent some kind of confidence intervals of the population loadings with a specific coverage percentage. All the convex hulls do is give the reader some visual impression of where the uncertainty in the PCA solution lies as a result of the missing data. A loading with a large convex hull is estimated with more uncertainty than a loading with a small convex hull, and the larger a convex hull is, the more cautious we must be regarding the interpretation of its loading. However, in order to assign some more absolute meaning to the convex hulls, Van Ginkel and Kroonenberg (2014) also studied what percentage of the $J \times K$ sample loadings that would be obtained if no data were missing is covered by the convex hulls under various circumstances. What they found was that under $M = 100$ imputations, the

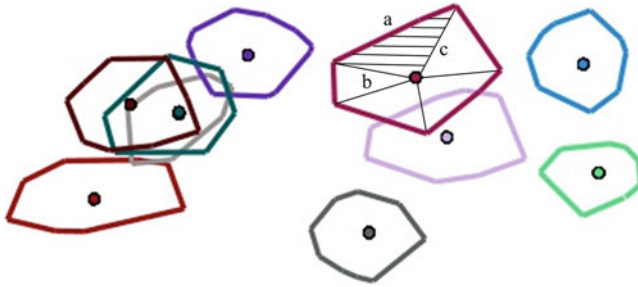


Fig. 8.2 Surface of one triangle in one of the convex hulls

convex hulls usually capture about 80% of the loadings that would be obtained if no data were missing, with percentages of missing data up to 15%. Based on these results, they gave a rough guideline to use $M = 100$ imputations if the researcher wants about 80% of the true sample loadings to fall within the corresponding convex hulls.

Finally, it should be noted that it is possible to use the convex hulls in the form of confidence intervals, using convex hull peeling (Green, 1981) or confidence ellipses (e.g., Josse et al., 2011). However, these confidence intervals do not make any statistical inference about a population loading, only about the true sample loading if no data were missing. A procedure for constructing confidence intervals of the population loadings will be discussed next.

8.5 Extensions

8.5.1 Confidence Intervals of the Component Loadings

As already mentioned in Sect. 8.4.2, in complete data there is the possibility of constructing confidence intervals of population component loadings. Analytical lower and upper bounds of confidence intervals have been derived by various authors (Girshick, 1939; Anderson, 1963; Archer & Jennrich, 1973; Ogasawara, 2000, 2002). However, these analytical confidence intervals have either been derived under the assumption that the data are multivariate normally distributed (Girshick, 1939; Anderson, 1963; Archer & Jennrich, 1973; Ogasawara, 2000), or they require a large sample size (Ogasawara, 2002).

Alternatively, bootstrap confidence intervals may be used for component loadings (Chatterjee, 1984; Efron & Tibshirani, 1994; Kiers, 2004; Lambert et al., 1990, 1991; Linting et al., 2007; Lorenza-Seva & Ferrando, 2003; Markus, 1994; Milan & Whittaker, 1995; Raykov & Little, 1999). Timmermans et al. (2007) studied two bootstrap procedures for component loadings in a simulation study, namely, the percentile method and the bias-corrected and accelerated (BCa) method (Efron,

1987). They found that both bootstrap procedures give better results regarding coverage of the component loadings than analytic methods.

Van Ginkel and Kiers (2011) proposed procedures to combine the bootstrap confidence intervals of both the percentile method and the BC_a method. Suppose that in the complete data case B , bootstrap samples are drawn for constructing confidence intervals of the loadings in \mathbf{A}_K , and the central $(1 - 2\alpha)$ part of the cumulative bootstrap distribution is the confidence interval. Van Ginkel and Kiers (2011) used the centroid solution $\mathbf{A}_{K,C}$ as the component matrix (see Sect. 8.4.1.2). Next, they drew B bootstrap samples from each of the M imputed datasets and used the central $(1 - 2\alpha)$ part of the total of $B \times M$ bootstrap samples as the confidence interval. They did this for both the percentile method and the BC_a method. In a simulation study, they investigated the statistical properties of their proposed procedures, and they turned out to produce coverage percentages close to the theoretical percentages, for various confidence widths (90%, 95%, and 99% coverage). The interested reader is referred to their paper.

8.5.2 Three-Mode Analysis

Three-mode analysis (e.g., Kroonenberg, 2008) is an extension of principal component analysis. It is used in datasets that consist of three different modes, for example, respondents (first mode) and questions on a questionnaire (second mode) at several different time points (third mode). The PCA model can be extended to a situation with three modes in several ways. The three most well-known extensions for three-mode data are the Tucker2 model (Tucker, 1972), the Tucker3 model (Tucker, 1966), and the Parafac model (Harshman, 1970; Carroll & Chang, 1970).

What all three models have in common is that they replace the singular value matrix $\mathbf{\Lambda}$ in Eq. (8.1) with a three-dimensional core array that also models the properties of the third mode, represented by different slices. Additionally, while in PCA $\mathbf{\Lambda}$ is always a square diagonal matrix, in the Tucker2 and Tucker3 model, the number of rows, columns, and slices of the core array are not necessarily the same. This implies that each mode (respondents, variables, time points) may be summarized by a different number of components. Furthermore, while the PCA model in Eq. (8.1) only has a matrix containing the scores of each respondent on the components (\mathbf{U}) and a matrix with scores of each variable on the components (\mathbf{V}), the Parafac and Tucker3 model also contain a matrix with scores of the third mode on the components.

Kroonenberg and Van Ginkel (2012) proposed rules for combining the results of the Tucker2 model in multiply imputed datasets. These combination rules are similar to the proposed combination rules discussed in Sect. 8.4.1.2. They involve applying generalized Procrustes analysis to both the three-mode equivalent of matrix \mathbf{U} and of matrix \mathbf{V} and by calculating the core matrix from both these two matrices and the M imputed datasets using matrix algebra. For the exact procedure, see Kroonenberg and Van Ginkel (2012).

Van Ginkel and Kroonenberg (2017) found that multiple imputation in combination with generalized Procrustes analysis produced good results of three-mode analysis in terms of RMSB (Eq. (8.2)) as compared to generalized least squares (see Sect. 8.3.2.2), the default method for handling missing data in three-mode analysis. It is, however, hard to tell what the specific influence of the combination techniques is on the RMSB, as for three-mode analysis there are no other combination techniques available than the one proposed by Kroonenberg and Van Ginkel (2012), to compare the procedure with.

8.6 Implementation in Software

Nowadays, most standard statistical software packages have included at least some procedure for creating multiply imputing incomplete datasets. Thus, when applying a PCA to an incomplete dataset, the question is not so much how to find a software package that can multiply impute the data as there are various options for that. The question is more which software package to use for combining the results of PCA on an incomplete dataset once it has been multiply imputed.

The software program `3WayPack` (The Three-Mode Company, 2021) is a freeware program that can be used for several three-mode models. The package also includes an option of using generalized Procrustes analysis. The program requires plain text as input, which is not really convenient when PCA results of multiply imputed datasets are printed in software specific output as they need to be converted to plain text first.

Alternatively, one can use the `shapes` package in R (Dryden & Mardia, 2016). This package can perform generalized Procrustes analysis. However, this package is more generally meant for the statistical analysis of landmark shapes and just happens to be also usable for combining results of PCA applied to multiply imputed dataset.

If one wants to stay completely within the framework of PCA on multiply imputed datasets, then the SPSS macro `GPA.sps` (Van Wingerde & Van Ginkel, 2021) may be used. This macro has been developed for applied researchers who use SPSS for their basic analyses and who want to combine the results of PCA within SPSS. The macro reads PCA output that has been saved to an SPSS data file, performs the calculations, and provides the (possibly Varimax rotated) matrix $\mathbf{A}_{K,C}$ in a new output. Plots with convex hulls as shown in Fig. 8.1 can also be printed.

8.7 Limitations and Final Considerations

Finally, a few limitations within the framework of PCA of multiply imputed datasets, and some points to take into consideration, will be discussed. As pointed out in this chapter, combination rules for component loadings in multiply imputed

datasets have been developed and investigated (e.g., Van Ginkel & Kiers, 2011; Van Ginkel & Kroonenberg, 2014; Van Ginkel et al., 2014). However, in PCA usually more outcomes are used and/or interpreted than only the component loadings.

For example, the component scores of the persons may need to be used for further analysis. At the moment not much has been written on how to compute component scores for multiply imputed datasets. Although not explicitly stated in their paper, Buisman et al. (2020) computed component scores for each imputed dataset m by standardizing the data to \mathbf{Z}_m and using $\mathbf{V}_m = \mathbf{Z}_m \sqrt{N} \mathbf{A}_{K,C}$. It has not been investigated, however, how this ad hoc solution performs in terms of bias in subsequent statistical analyses with these component scores.

As a second example, no combination rules have been defined for the proportion of variance accounted for by the extracted components. One could construct a pooled $\mathbf{\Lambda}$ matrix using a similar procedure for constructing the core three-way array in three-mode analysis discussed in Sect. 8.5.2 (also, see Kroonenberg & Van Ginkel, 2012). Next, the first K singular values of the pooled $\mathbf{\Lambda}$ could be used for getting a measure for the total amount of explained variance. At present the theoretical properties of such a solution have not been derived nor investigated. Consequently, it is currently unknown how closely such an estimate of the proportion of explained variance resembles the proportion of explained variance that would have been obtained if the data had been complete.

In short, there are still things that remain to be developed and investigated regarding the pooling of estimates and statistics within PCA applied to multiply imputed data. This is more generally a problem of multiple imputation. Rubin (1987) provided only very general combination rules for statistical analyses that can be applied when a parameter estimate or a set of parameter estimates is tested for significance. For some statistics and analyses that do not directly fit into that framework, additional combination rules have been developed since Rubin (1987), but for other statistics and analyses, there is still work to be done regarding combination rules. Whenever applied researchers are interested in statistics or analyses for which no combination rules are available yet, they are often inclined to set aside multiple imputation as a method for handling their missing data altogether.

However, Van Ginkel et al. (2020) argue that even when combination rules for specific analyses and statistics are lacking, it may not always be harmful to use something ad hoc. Even without a theoretical justification, ad hoc solutions can still give a rough but reasonable indication of what the actual statistic would have been without missing data. Additionally, since PCA is usually (but not always) used without any statistical testing, one cannot draw erroneous conclusions as a result of type I or type II errors. Even when something as simple as averaging $\mathbf{\Lambda}$'s across imputed datasets is done, this will probably still give a good indication of how many of the components contribute substantially to the explained total variance and which do not.

In summary, when a PCA needs to be carried out on an incomplete dataset, multiple imputation may be a good tool to handle the missing data. Although pairwise deletion does not necessarily give worse results than multiple imputation, multiple imputation comes with many other advantages, such as all analyses being

applied to the dataset being comparable regarding sample size and cases being included in the analyses. Besides, pairwise deletion has the disadvantage that computational problems may occur. Estimates of component loadings in multiply imputed datasets can readily be computed using generalized Procrustes analysis. Other statistics in PCA may not have combination rules as of yet, but using some quick-and-dirty procedures may not be harmful for the given purposes of PCA.

References

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1), 122–148. <http://www.jstor.org/stable/2991288>
- Archer, C. O., & Jennrich, R. I. (1973). Standard errors for rotated factor loadings. *Psychometrika*, 38(4), 581–592. <https://doi.org/10.1007/BF02291496>
- Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and em methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate behavioral research*, 35(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03
- Borgonovi, F., & Pál, J. (2016). *A framework for the analysis of student well-being in the pisa 2015 study*. 140. <https://doi.org/10.1787/5jlpszwghvnb-en>
- Boulton, M. J., Bucci, E., & Hawker, D. D. (1999). Swedish and English secondary school pupils' attitudes towards, and conceptions of, bullying: Concurrent links with bully/victim involvement. *Scandinavian Journal of Psychology*, 40(4), 277–284. <https://doi.org/10.1111/1467-9450.404127>
- Buisman, R., Pittner, K., Tollenaar, M. S., Lindenberg, J., van den Berg, L., Compier-de Block, L., van Ginkel, J. R., Alink, L., Bakermans-Kranenburg, M. J., Elzinga, B. M., & van IJzendoorn, M. H. (2020). Intergenerational transmission of child maltreatment using a multi-informant multi-generation family design. *PLoS One*, 15(3), e0225839. <https://doi.org/10.1371/journal.pone.0225839>
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3), 283–319. <https://doi.org/10.1007/BF02310791>
- Chatterjee, S. (1984). Variance estimation in factor analysis: An application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, 37(2), 252–262. <https://doi.org/10.1111/j.2044-8317.1984.tb00803.x>
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Psychology Press. <https://doi.org/10.4324/9781315827506>
- 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)*, 39(1), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Dryden, I. L., & Mardia, K. V. (2016). *Statistical shape analysis: With applications in R*. Wiley.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185. <https://doi.org/10.2307/2289144>
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429246593>
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507–521. <https://doi.org/10.2307/2331838>
- Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Sage.
- Girshick, M. A. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics*, 10(3), 203–224. <https://doi.org/10.1214/aoms/1177732180>

- Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the archaeological & historical sciences* (pp. 138–149). Edinburgh University Press.
- Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51. <https://doi.org/10.1007/bf02291478>
- Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, *17*, 429–440. <https://doi.org/10.1007/BF02288918>
- Green, P. J. (1981). Peeling bivariate data. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 3–19). Wiley.
- Grung, B., & Manne, R. (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *42*(1), 125–139. [https://doi.org/10.1016/S0169-7439\(98\)00031-8](https://doi.org/10.1016/S0169-7439(98)00031-8)
- Harman, H. H. (1976). *Modern factor analysis*. University of Chicago Press.
- Harshman, R. A. (1970). Foundation of the PARAFAC procedure: Models and condition for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1–84.
- Hills, P., Francis, L. J., & Robbins, M. (2005). The development of the Revised Religious Life Inventory (RLI-R) by exploratory and confirmatory factor analysis. *Personality and Individual Differences*, *38*(6), 1389–1399. <https://doi.org/10.1016/j.paid.2004.09.006>
- Josse, J., Husson, F., & Pagès, J. (2009). Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la société française de statistique*, *150*(2), 28–51. http://www.numdam.org/item/JSFS_2009__150_2_28_0/
- Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, *5*(3), 231–246. <https://doi.org/10.1007/s11634-011-0086-7>
- Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, *62*(2), 251–266. <https://doi.org/10.1007/BF02295279>
- Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, *18*(1), 22–36. <https://doi.org/10.1002/cem.841>
- Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Wiley.
- Kroonenberg, P. M., & van Ginkel, J. R. (2012). Combination rules for multiple imputation in three-way analysis illustrated with chromatography data. *Current Analytical Chemistry*, *8*(2), 224–235. <https://doi.org/10.2174/157341112800392544>
- Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement*, *50*(1), 33–48. <https://doi.org/10.1177/0013164490501004>
- Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, *26*, 421–434.
- Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, *12*(3), 359–379. <https://doi.org/10.1037/1082-989X.12.3.359>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119013563>
- Lorenza-Seva, U., & Ferrando, P. J. (2003). IMINCE: An unrestricted factor-analysis-based program for assessing measurement invariance. *Behavior Research Methods, Instruments, and Computers*, *35*, 318–321.
- Markus, M. T. (1994). Bootstrap confidence regions for homogeneity analysis; the influence of rotation on coverage percentages. In R. Dutter & W. Grossmann (Eds.), *Compstat*. Physica. https://doi.org/10.1007/978-3-642-52463-9_38
- Mata, I., Mataix-Cols, D., & Peralta, V. (2005). Schizotypal personality questionnaire-brief: Factor structure and influence of sex and age in a nonclinical population. *Personality and Individual Differences*, *38*(5), 1183–1192. <https://doi.org/10.1016/j.paid.2004.08.001>
- Meulman, J. J. (1982). *Homogeneity analysis of incomplete data*. DSWO Press.

- Milan, L., & Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(1), 31–49. <https://doi.org/10.2307/2986193>
- Ogasawara, H. (2000). Standard errors of the principal component loadings for unstandardized and standardized variables. *The British Journal of Mathematical and Statistical Psychology*, 53(2), 155–174. <https://doi.org/10.1348/000711000159277>
- Ogasawara, H. (2002). Concise formulas for the standard errors of component loading estimates. *Psychometrika*, 67(2), 289–297. <https://doi.org/10.1007/BF02294847>
- Raykov, T., & Little, T. D. (1999). A note on procrustean rotation in exploratory factor analysis: A computer intensive approach to goodness-of-fit evaluation. *Educational and Psychological Measurement*, 59(1), 47–57. <https://doi.org/10.1177/0013164499591004>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Pearson.
- Takane, Y., & Oshima-Takane, Y. (2003). Relationships between two methods for dealing with missing data in principal component analysis. *Behaviormetrika*, 30(2), 145–154. <https://doi.org/10.2333/bhmk.30.145>
- Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2), 267–276. <https://doi.org/10.1007/BF02294053>
- The Three-Mode Company Home Page. (2021). Retrieved 15 October 2021, from <https://three-mode.leidenuniv.nl/>
- Timmerman, M. E., Kiers, H. A. L., & Smilde, A. K. (2007). Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, 60(2), 295–314. <https://doi.org/10.1348/000711006X109636>
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3), 279–311. <https://doi.org/10.1007/BF02289464>
- Tucker, L. R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika*, 37(1, Pt. 1), 3–27. <https://doi.org/10.1007/BF02291410>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429492259>
- Van Ginkel, J. R., & Kiers, H. A. L. (2011). Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: A multiple-imputation approach. *British Journal of Mathematical and Statistical Psychology*, 64(3), 498–515. <https://doi.org/10.1111/j.2044-8317.2010.02006.x>
- Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Using generalized procrustes analysis for multiple imputation in principal component analysis. *Journal of Classification*, 31(2), 242–269. <https://doi.org/10.1007/s00357-014-9154-y>
- Van Ginkel, J. R., & Kroonenberg, P. M. (2017). Evaluation of multiple-imputation procedures for three-mode component models. *Journal of Statistical Computation and Simulation*, 87(16), 3059–3081. <https://doi.org/10.1080/00949655.2017.1355368>
- Van Ginkel, J. R., Kroonenberg, P. M., & Kiers, H. A. L. (2014). Missing data in principal component analysis of questionnaire data: A comparison of methods. *Journal of Statistical Computation and Simulation*, 84(11), 2298–2315. <https://doi.org/10.1080/00949655.2013.788654>
- Van Ginkel, J. R., Linting, M., Rippe, R., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 102(3), 297–308. <https://doi.org/10.1080/00223891.2018.1530680>
- Van Wingerde, B., & van Ginkel, J. (2021). SPSS syntax for combining results of principal component analysis of multiply imputed data sets using generalized procrustes analysis. *Applied Psychological Measurement*, 45(3), 231–232. <https://doi.org/10.1177/0146621621990757>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

