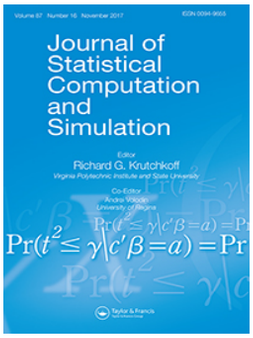

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Evaluation of multiple-imputation procedures for three-mode component models

Joost R. van Ginkel & Pieter M. Kroonenberg

To cite this article: Joost R. van Ginkel & Pieter M. Kroonenberg (2017) Evaluation of multiple-imputation procedures for three-mode component models, Journal of Statistical Computation and Simulation, 87:16, 3059-3081, DOI: 10.1080/00949655.2017.1355368

To link to this article: http://dx.doi.org/10.1080/00949655.2017.1355368

Published online: 25 Jul 2017.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2017
VOL. 87, NO. 16, 3059–3081
https://doi.org/10.1080/00949655.2017.1355368

Taylor & Francis
Taylor & Francis Group

Evaluation of multiple-imputation procedures for three-mode component models

Joost R. van Ginkel and Pieter M. Kroonenberg

Child and Family Studies, Leiden University, Leiden, the Netherlands

ABSTRACT

Three-mode analysis is a generalization of principal component analysis to three-mode data. While two-mode data consist of cases that are measured on several variables, three-mode data consist of cases that are measured on several variables at several occasions. As any other statistical technique, the results of three-mode analysis may be influenced by missing data. Three-mode software packages generally use the expectation–maximization (EM) algorithm for dealing with missing data. However, there are situations in which the EM algorithm is expected to break down. Alternatively, multiple imputation may be used for dealing with missing data. In this study we investigated the influence of eight different multiple-imputation methods on the results of three-mode analysis, more specifically, a Tucker2 analysis, and compared the results with those of the EM algorithm. Results of the simulations show that multilevel imputation with the mode with the most levels nested within cases and the mode with the least levels represented as variables gives the best results for a Tucker2 analysis. Thus, this may be a good alternative for the EM algorithm in handling missing data in a Tucker2 analysis.

ARTICLE HISTORY
Received 24 June 2016
Accepted 11 July 2017

KEYWORDS
Missing data; multiple imputation; multilevel; three-mode analysis; Tucker2 model

1. Introduction

Three-mode analysis [1–3] is an extension of principal component analysis (PCA) to three-mode data. While two-mode data consist of cases measured on several variables, three-mode data consist of cases (first mode) measured on the same variables (second mode) at several occasions, or conditions (third mode). In a simulation study, we will compare procedures for missing data in the context of three-mode analysis. Specifically, we will concentrate on the Tucker2 model; in the discussion, we will briefly comment on other three-mode models.

1.1. Missing data

Before getting into the technical details of three-mode analysis, the problem of missing data and its solutions are explained. As in any other statistical technique, missing data may complicate a three-mode analysis. Many statistical techniques are designed to only handle

CONTACT Joost R. van Ginkel jginkel@fsw.leidenuniv.nl Child and Family Studies, Leiden University, Wassenaarseweg 52, Leiden AK 2333, the Netherlands

complete data, so that cases with missing data are automatically excluded by the statistical software package. This is called listwise deletion. Besides being very wasteful, listwise deletion is only guaranteed to give unbiased results for the parameters when the missing data are a random subsample of all data points. When this condition holds, the missing data are said to be *missing completely at random* (MCAR). For a more technical explanation of MCAR, see Little and Rubin [4, p. 10]. Unfortunately, MCAR almost never holds, so that listwise deletion is generally not a good option.

The assumption of MCAR may be relaxed by assuming that the missing data depend on observed variables but not on unobserved variables. For example, consider a data set in which age is observed for all respondents, but as age increases, the probability of missing data on other variables increases as well. This is called *missing at random* (MAR [4, p. 10], [5]). Under MAR, listwise deletion is not guaranteed to give unbiased results.

However, three-mode models already have a facility for handling missing data, namely the expectation–maximization (EM) algorithm [6,7]. The EM algorithm alternates between estimating the parameters of a statistical model of interest (here, a three-mode model) and estimating the missing data based on the current parameter estimates. Besides using all cases for the analysis (contrary to listwise deletion), this technique will also give unbiased results under specific types of MAR, namely when the missing data depend on observed variables included in the statistical model of interest. However, when missing data depend on variables not included in the model, the MAR assumption is violated for the specific analysis, and the data are called *not MAR* [4, p. 10]. Under the more general situation of not missing at random (NMAR), missingness depends on unobserved variables or information. Whenever a variable explaining the missingness is not included in the statistical model of interest, it is equivalent to missingness depending on an unobserved variable because it does not take part in handling the missing data.

Alternatively, one could use *multiple imputation* [8]. This procedure works in three steps: (1) the multiple-imputation step: missing data are estimated multiple times using an estimation method, from here on denoted as *multiple-imputation method*, so that multiple complete versions of the incomplete data set are created; (2) the *multiple-analysis step*: the completed data sets are analysed using the statistical technique of interest; (3) the *pooling step*: the results of the separate analyses are pooled into one overall analysis, taking into account the additional uncertainty due to the missing data in the standard errors and p-values.

Unlike the EM algorithm, multiple imputation can include variables outside the statistical model of interest for explaining the missingness. As long as the variables explaining the missingness are observed and used by multiple imputation for imputing the data, the MAR assumption is not violated, whether or not they take part in the subsequent statistical analysis (here, a three-mode analysis).

1.2. Two-mode PCA

Since three-mode analysis is a generalization of PCA to three-mode data, the technique is best explained by briefly discussing PCA first. In PCA, a large number of variables J are summarized into a smaller number of components, S. It is commonly applied to standardized data $\mathbf{Z}$ (size $I \times J$), especially when variables are measured on different scales. PCA

decomposes **Z** as

$$\mathbf{Z} = \mathbf{A}\Lambda\mathbf{B}' + \mathbf{E}. \tag{1}$$

Here, **A** is an $I \times S$ matrix of coefficients of all cases on the first S components, obtained from a singular value decomposition of **Z**. Secondly, $\mathbf{\Lambda}$ is a diagonal $S \times S$ matrix of the first S singular values of the singular value decomposition. Thirdly, **B** is a $J \times S$ matrix of the variable coefficients on the first S components. Finally, **E** is an $I \times J$ matrix with random errors. The PCA model can be written as follows:

$$\mathbf{Z} = \mathbf{A}\mathbf{F}' + \mathbf{E}, \tag{2}$$

where $\mathbf{F} = \mathbf{B}\mathbf{\Lambda}$ is a $J \times S$ matrix with standardized coefficients.

1.3. Three-mode PCA: Tucker2 model

The above model has several three-mode generalizations, in particular, the Tucker2 model [9], the Tucker3 model [3], and the Parafac model [10,11]. Although the Tucker2 model is used less frequently than the Parafac and Tucker3 model, the focus will be on the Tucker2 model, for two reasons. Firstly, this study further elaborates on a paper about three-mode analysis and multiple imputation [12]. Kroonenberg and Van Ginkel discussed rules for combining the results of a Tucker2 analysis to a multiply imputed dataset. Secondly, the Tucker2 model has fewer component matrices than the Tucker3 and Parafac model, which in a simulation study allows for a substantially more focused discussion of the results than for the Tucker3 and Parafac model. The Tucker3 and Parafac models will be commented in Section 5.

Suppose we have an I (cases) $\times$ J (variables) $\times$ K (occasions) three-mode data set $\mathrm{Z} = (z_{ijk})$ in which $\mathrm{Z}_k = (z_{ij}^k)$ is the kth slice of the third mode. Next, let P be the number of components for the first mode (cases) and Q the number of components of the second mode (variables), Q not necessarily equal to P. Finally, let **H** be a $P \times Q \times K$ three-mode core array consisting of K, not necessarily diagonal, $P \times Q$ slices $\mathbf{H}_k$. The Tucker2 model is defined as

$$\mathbf{Z}_k = \mathbf{A}\mathbf{H}_k\mathbf{B}' + \mathbf{E}_k \quad (k = 1, \ldots, K). \tag{3}$$

Thus, the three-mode data are modelled for each level k of the third mode (occasion) by common subject components, common variable components and a level-specific core slice which contains the strengths of the links between the two types of components.

1.4. Multiple imputation and three-mode analysis

Although multiple imputation is a good solution for missing data in many statistical analyses, in applying multiple imputation prior to a three-mode analysis one is faced with two challenges. The first challenge is to find a multiple-imputation method that preserves the results of the three-mode analysis as much as possible. The second is the pooling of the parameter estimates of the three-mode model into one set of parameters. Both challenges are discussed next.

1.4.1. Multiple-imputation methods for three-mode data

In general, two basic types of multiple-imputation procedures exist: *joint modelling* (see Van Buuren [13, p. 105–108]) and *fully conditional specification* [13, p. 118–116]. The latter can be further subdivided into the *regression* and the *predictive mean matching* approach.

1.4.1.1. Joint modelling. In joint modelling, one joint distribution of the data is assumed, and used for imputation. For example, Schafer [14, Chapter 5] discusses multiple imputation under a multivariate normal distribution. Joint modelling approaches use an algorithm called *data augmentation* [15]. This algorithm iteratively draws random values of the unknown parameters of this multivariate normal distribution (i.e. population means and the covariance matrix) from a posterior distribution given the observed data, and then randomly draws the imputed values from the distribution with the obtained parameters of the model. The variance of the resulting multiply imputed values reflects both uncertainty about the unknown parameters of the joint distribution, and uncertainty about the missing data.

1.4.1.2. Fully conditional specification. Like joint modelling, fully conditional specification takes into account uncertainty about both the unknown parameters and the unknown values of the missing data, but works differently. Rather than using one joint distribution for imputing the data, this method uses a conditional distribution for each variable separately. That is, for imputing the data it uses a conditional distribution based on normal linear regression (in case of a numerical variable) or multinomial logistic regression (in case of a categorical variable) with the other variables as predictors. The underlying algorithm first fills in starting values for the missing data, next it iterates across all variables where for each iteration it randomly draws the parameters of the regression model from a posterior distribution. Next, it randomly draws the imputed values from the conditional distribution based on the (multinomial logistic) regression model, given the drawn parameters. On top of the iterations across variables, it iteratively repeats the process until properties of the imputed values (e.g. means and standard deviations), stabilize. For technical details, see Van Buuren [13], Van Buuren et al. [16,17].

An advantage of fully conditional specification over joint modelling is that it is more flexible. While in joint modelling, all variables in the imputation model are used for imputation of missing data on all other variables, fully conditional specification can use different predictors for each variable with missing data. This is especially useful when the data set contains many variables. By using only predictors that may be relevant for estimating the missing data on a specific variable, overfitting of the imputation model can be avoided.

1.4.1.3. Predictive mean matching. Besides the standard fully conditional specification procedure, another variant exists, namely *predictive mean matching* (PMM; Van Buuren [13, p. 68–74], Van Buuren et al. [16,17], and Rubin [18]). In predictive mean matching, a linear regression model is used for imputing the missing data on continuous variables as well, but here the imputed values are not drawn from the conditional distribution based on the regression model. Instead, the regression model is used for finding respondents with observed values on the outcome variable whose predicted values on the outcome variable closely resemble the predicted values of the respondents with missing values. For each person with a missing value on a particular variable, the observed value of the matching

respondent is used for imputation. This variant has been shown to be more robust to model violations than standard fully conditional specification [19,20]. Both versions of fully conditional specification are implemented in SPSS 23.0 [21], the mice package in R [22], and in SAS 12.1 [23].

A problem with applying the above procedures to three-mode data is that they were designed for two-mode data and that consequently, the data have to be rearranged first, such that they can be conceived of as a two-mode data set. A number of possible ways to do this are discussed below.

1.4.1.4. Long imputation. The first option for rearranging a three-mode data set to a two-mode data set is in long format: occasions of the same case are represented in separate rows as if they were independent cases. Next, the data are imputed using the standard multiple-imputation methods. A problem of this approach is that it may bias the dependence among occasions towards independence. For the three-mode model that is applied next, this implies that the first mode (cases) and third mode (occasions) are treated as one mode in the multiple-imputation method, so that the data get biased towards a two-mode structure of *persons/occasions* × *variables*.

1.4.1.5. Wide imputation. The second option is storing the data in wide format: variables at each occasion are represented by different columns, as if the variables measured on a new occasion are completely new variables. Next, data are imputed using the standard multiple-imputation methods. In this way, dependence among time points is maintained. However, this way of imputing ignores the fact that the data have a hierarchical structure in which the same variables are measured at several occasions. For the three-mode model that is applied to this data set this means that the data get biased towards a two-mode structure of *persons* × *occasions/variables*. Another disadvantage compared to long imputation is that when a dataset is restructured into wide format it has more variables and fewer cases than in long format, which may increase the risk of overfitting.

1.4.1.6. Imputation for separate slices. A third option is to treat each slice as a two-mode data set and impute each slice separately using the standard multiple-imputation methods. A disadvantage of this approach is that it biases the data towards a structure where both cases and variables have no common variance over occasions. For the three-mode model that is applied next, this implies that the data get biased towards a structure without common matrices **A** and **B**, but k separate **A**s and **B**s, i.e. the data get biased towards a structure of k separate PCA models. For the three-mode analysis that is carried out next this could mean that this bias towards independent slices is directed to the core array, which will consequently be biased as well.

1.4.1.7. Multilevel imputation. Conceptually, the best way to carry out multiple imputation for three-mode data, is to use a multiple-imputation method that explicitly models a three-mode structure, such as multiple imputation under a multilevel model [24]. More generally, multilevel analysis is used for modelling data with a hierarchical structure, for example, data consisting of students nested within classes. A special application of multilevel is the modelling of longitudinal data where different time points are nested within cases [24, Chapter 4]. Three-mode data have the structure of a multilevel data set in the

sense that different occasions (third mode) are nested within cases (first mode). Consequently, multilevel imputation is an appropriate way to impute missing data in three-mode data.

Currently, there are two approaches to perform multiple imputation under a multilevel model, namely a joint modelling variant [25], and a fully conditional specification variant [26]. Unfortunately, no multilevel predictive mean matching variant has been proposed so far. The joint modelling variant has been implemented in the pan [27] and mice [22] procedures in R; mice also contains a fully conditional specification variant.

1.4.1.8. Multiple-imputation methods in this study. When missing data are handled using a multiple-imputation method not based on a multilevel model, a number of decisions must be made about the procedure. First a decision must be made about how to apply a method for two-way data to a three-way data set. As mentioned before, the data can be put in long format (L), in wide format (W), or the imputation procedure may be applied to the separate slices (S). Secondly, a decision must be made about whether to use the regression approach (Reg) or the PMM approach. Using all combinations of ways to apply two-mode multiple-imputation methods to three-mode data (L, W, S) and estimation procedures (Reg, PMM), this results in six two-mode multiple-imputation procedures adapted for three-mode data, denoted by Reg-L, Reg-W, Reg-S, PMM-L, PMM-W, and PMM-S.

To be consistent with the other imputation methods, it was initially decided to use the fully conditional specification variant of multilevel imputation. However, this version ran into computational problems for large percentages of missingness (i.e. imputed values far beyond the range of the data), so it was necessary to switch to joint modelling [27].

The multilevel approach may be used in two ways. The first way is to use variables measured at different occasions as model variables, and to nest different occasions within cases. This method will be denoted PAN. The second way is to use different occasions as model variables, and to nest variables measured at different occasions within cases. This method will be denoted PAN-restructured (PAN-R). Both options were studied.

Although conceptually it makes more sense to use variables as model variables and nest occasions within cases (PAN), it may still be beneficial to do it the other way around (PAN-R). When $J > K$, time points are modelled as variables, and variables are nested within cases, there will be relatively few model variables in the imputation model while at the same time more information can be used for estimating the random effect. This way of modelling the hierarchical structure will reduce the risk of sparse data.

As an aside, one must bear in mind that an imputation model differs from an analysis model in its purpose. The latter is used for drawing inferences about the parameters of the analysis model while the former is only used to get imputed values that closely mimic the structure of the observed data [17, p. 143].

Together with the default EM option in three-mode analysis, this resulted in nine different missing-data methods in our study, which are summarized in Table 1.

1.4.2. Pooling the results of three-mode analysis

After completing the data in the multiple-imputation step, the second (multiple-)analysis step consists of applying statistical analyses to the multiple completed data sets. In the third or pooling step, the results of these multiple analyses are combined to one pooled result. For significance tests based on z – or t-tests, combination rules have been defined in [8] and

Table 1. Overview of the missing-data methods included in the simulation study.

Joint modelling/fully conditions specification	Regression/Predictive mean matching/Multilevel	Data format EM algorithm	Name EM
Fully conditional specification	Regression	Long	Reg-L
		Wide	Reg-W
		Separate slices	Reg-S
	Predictive mean matching	Long	PMM-L
		Wide	PMM-W
		Separate slices	PMM-S
Joint Modelling	Multilevel	Long (occasions within cases)	PAN
		Long (variables within crabs)	PAN-R

are available in most statistical software packages. Pooling techniques suitable for testing multiple parameters at a time (e.g. an overall F-test) have also been available for a long time [8], but were only recently explicitly worked out for analysis of variance [28].

When no significance tests are desired, it is sufficient to simply average over the multiple values of a statistic derived from the multiply imputed data sets. However, Van Ginkel and Kroonenberg [29] argued and showed that for component loadings in PCA this is not adequate. The problem with averaging PCA loadings is that firstly, in one imputed data set the loadings of a specific component may have opposite signs compared to the loadings of the same component in another imputed data set. Secondly, the order of the components may not be the same for all imputed data sets. Rather than averaging loadings, Van Ginkel and Kroonenberg showed that one should use Generalized Procrustes analysis (GPA; [30,31]) for combining the PCA loadings from multiply imputed data sets. GPA is a generalization of a Procrustes rotation [32,33]. In a Procrustes rotation, a source matrix is rotated towards a target matrix, such that it optimally fits the target. In GPA this idea is extended to more than two matrices. Consequently, in GPA the distinction between source and target matrix gets lost. Instead, all matrices are optimally aligned towards each other, and the centroid of the optimally aligned matrices may serve as an average of all matrices. In the context of multiple imputation, different PCA solutions of multiply imputed datasets are optimally aligned using GPA, and the centroid is the pooled PCA solution.

The same approach may be used for three-mode analysis. Kroonenberg and Van Ginkel [12] defined combination rules for the Tucker2 model, based on the GPA approach for PCA. In their proposed procedure, pooled component matrices **A** and **B** are obtained first, using the GPA approach for both matrices. Next, **H** is constructed from **A** and **B**. To explain the pooling of **H**, we first consider the case where no data are missing. When the data are complete, matrix (slice) $\mathbf{H}_k$ is computed as follows:

$$\mathbf{H}_k = \mathbf{A}'\mathbf{Z}_k\mathbf{B} \tag{4}$$

after first computing matrices **A** and **B** using an algorithm were described by Kroonenberg and De Leeuw [2]. For multiply imputed data, $\mathbf{H}_k$ is computed for each imputed data set m as

$$\mathbf{H}_{k,m} = \mathbf{A}'\mathbf{Z}_{k,m}\mathbf{B} \tag{5}$$

using the solutions of **A** and **B** obtained from GPA. Next, $\mathbf{H}_k$ is obtained by averaging the $\mathbf{H}_{k,m}$'s over the M imputed data sets.

1.4.3. Research questions

Besides Kroonenberg and Van Ginkel [12], other studies about missing data in three-mode data include Hubert et al. [34], Louwerse et al. [35], and Tian et al. [36]. The paper by Kroonenberg and Van Ginkel [12] was aimed at defining combination rules for the Tucker2 model for multiply imputed data sets. However, the problem of accurately taking the three-mode structure into account in the imputation model was not addressed (wide imputation was used) and no systematic simulation study was conducted.

Hubert et al. [34] discussed an EM-based estimation method for the Parafac model that was robust to both outliers and incomplete data. Outliers are not the topic of the present paper so the only EM-based method in our study will be the standard EM algorithm.

In [35] the focus was on cross-validation methods for three-mode models in the presence of missing data. In the current study, we will not focus on ways to select the best three-mode model for an incomplete data set, but on how parameter estimates of a three-mode model are influenced by multiple imputation, assuming the best model has already been selected.

Finally, Tian et al. [36] looked at the performance of multiple imputation in three-mode data at the level of the imputed data themselves, but the influence on the parameters of a three-mode model was not studied. Moreover, Tian et al. only looked at multiple-imputation methods for two-way data, adjusted such that they could be applied to three-mode data. In the current study, we have focused on the influence of multiple imputation on the parameter estimates of a Tucker2 analysis. In doing so, we will both look at multiple-imputation methods for two-way data applied to a three-mode data set, and multiple-imputation methods that explicitly model the three-mode structure of the data. This leads us to the following research questions: (1) Is there any benefit of multiple imputation over the EM algorithm in a Tucker2 analysis and if so, (2) is it sufficient to use the standard established multiple-imputation methods for two-way data with some adjustments, or is a multiple-imputation method that explicitly models the three-mode structure necessary?

2. Method

For our simulation study, the following procedure was adopted: Complete data sets were simulated using a Tucker2 model. These complete data sets were made incomplete by removing various amounts of data points according to different stochastic mechanisms, to be explained later. For each of these incomplete data sets, a Tucker2 model was estimated using either multiple-imputation methods outlined above, or the EM algorithm, and a quality measure was computed for each solution.

2.1. Creating the simulated data

2.1.1. Simulating the complete data

To simulate data sets according to a Tucker2 model, the parameter values were taken from a Tucker2 model, estimated from an existing data set, namely the Blue Crab data [37,38]. The data set consists of $N = 138$, possibly diseased crabs ($n = 48$ in 1989, and $n = 90$ in 1990) in two different regions, namely the Pamlico River, and Albemarle Sound, both in North Carolina. Within the Pamlico River, two types of crabs were to be found, namely healthy

Table 2. Estimates of the **B** and **H** matrix of the Tucker2 model applied to the Crab data [37].

		Component number second way			
Matrix	Trace element	1	2	3	4
B	Al	0.45	0.00	−0.15	−0.12
	Ca	−0.01	0.53	0.06	−0.09
	Cd	0.03	−0.42	0.16	−0.15
	Cr	0.45	0.01	−0.04	0.03
	Cu	−0.18	−0.09	−0.10	0.29
	K	−0.01	0.00	0.43	0.23
	Mg	0.07	0.32	0.54	0.02
	Mn	0.33	0.06	0.04	−0.16
	Na	−0.02	−0.04	0.54	−0.43
	Ni	0.31	−0.10	0.24	0.23
	P	−0.02	0.54	−0.07	0.07
	Pb	0.19	−0.03	0.17	0.70
	Se	−0.07	−0.26	0.03	0.15
	Ti	0.43	0.00	−0.06	−0.02
	U	0.34	−0.04	−0.15	−0.12
	Zn	0.02	−0.24	0.22	−0.14
Component number first way					
$\mathbf{H}_{\text{Gill}}$	1	42.52	0.93	−0.35	2.67
	2	6.41	2.66	2.11	−14.12
	3	1.94	−6.20	−19.71	−4.74
$\mathbf{H}_{\text{Hepatopancreas}}$	1	1.15	9.02	2.74	−0.97
	2	2.00	−29.45	6.01	−0.67
	3	−1.33	−1.94	−11.58	1.66
$\mathbf{H}_{\text{Muscle}}$	1	0.77	1.50	1.10	−2.08
	2	0.25	5.09	0.85	−0.46
	3	−1.33	0.61	−5.94	5.72

crabs and diseased crabs. Thus, three different categories of crabs were available, namely Albemarle Sound, healthy ($n = 46$), Pamlico River, healthy ($n = 46$), and Pamlico River, diseased ($n = 46$). This grouping variable was used as a background/external variable for simulating a MAR mechanism, as will be explained later.

In the original study by Gemperline et al. [37], it was investigated whether trace-element levels were associated with the occurrence of the disease. In total, three tissue types (gill, hepatopancreas, and muscles), were sampled from each crab, and the levels of 25 trace elements in these tissues were determined. In our study we used the 16 most abundant trace elements: Al, Ca, Cd. Cr, Cu, K, Mg, Mn, Na, Ni, P, Pb, Se, Ti, U, and Zn [38], so that the resulting simulated data sets had a 138 (crabs) × 16 (trace elements) × 3 (tissue types) three-mode data structure.

Because the measurement units of the trace elements were highly different, the data were centred and normalized. In particular, first each trace-element tissue-type combination (jk) was centred separately across crabs, i.e. the data were centred by $\bar{z}_{.jk}$ [1, p. 119–121]. Next, each centred trace element (j) was normalized across all crabs and tissue types, i.e. they were normalized by $s_{.j}$; see, e.g. Kroonenberg [1, p. 125–127].

Using deviance and multiway scree plots (Kroonenberg [1, Chapter. 8]; Timmerman and Kiers [39]), the best-fitting Tucker2 model was chosen, which turned out to be a 3 (crabs) × 4 (trace elements) Tucker2 model. The estimates of **B** and **H** (see Equation (6)) were used in the process of simulating the data (see Table 2).

Applying multiple imputation to three-mode data only makes sense when cases are random [1, p. 168]. Therefore, we cannot use the scores matrix **A** from the Crab data

Table 3. Means and covariance matrices of matrix **A** of the Tucker2 model applied to the Crab data (Gemperline et al. [37]), for each group of crabs separately.

		Covariance matrices			
		Component number first way			
	Component number first way	1	2	3	Mean
$\mathbf{A}_{\text{Alblemare}}$	1	11			−5411
	2	−17	264		−3073
	3	−27	−69	355	4438
$\mathbf{A}_{\text{Pamlico}-\text{h}}$	1	205			−1589
	2	−410	1180		495
	3	−212	29	318	−5522
$\mathbf{A}_{\text{Pamlico}-\text{d}}$	1	1181			7001
	2	80	612		2578
	3	318	−80	714	1084

Note: Entries have been multiplied by 10^5.

for simulating the data because a fixed **A** would imply that the cases (crabs) are fixed. Instead, matrix **A** obtained from the Tucker2 analysis of the Blue Crab data was divided into three sub-matrices, one for each of the three group of Crabs. The group means and group covariance matrices of each of these sub-matrices (see Table 3) were used to construct sub-matrices randomly drawn from a multivariate normal distribution with the same means and covariance matrices. Together these sub-matrices formed one overall simulated **A** matrix.

Although it is questionable whether a multivariate normal distribution of matrix **A** is realistic, the multivariate normal distribution was chosen to be consistent with the distributional assumptions of most of the multiple-imputation methods used. One of the purposes of this study was to see whether multiple imputation for two-way data could already accurately recover the results of a Tucker2 analysis or if explicit modelling of the three-mode structure in the multiple-imputation process would be necessary, regardless of possible violations of distributional assumptions. Studying the robustness of these methods to violations of multivariate normality is not the topic of this paper.

One requirement of matrices **A** and **B** in the Tucker2 model is that they are both orthonormal. While this is true for the matrix **A** from the analysis of the original Blue Crab data, the sampled **A** matrices are not orthonormal due to sampling fluctuation. Consequently, when the sampled **A** is inserted in Equation (3), the resulting data set **Z** is not a data set that exactly behaves according to the Tucker2 model with the parameters from Tables 1 and 2. To create a sampled orthonormal matrix $\mathbf{A}_\nu$ of simulated data set ν, the originally drawn sub-matrices $\mathbf{A}_{\text{Alblemare},\nu}$, $\mathbf{A}_{\text{Pamlico}-\text{h},\nu}$, and $\mathbf{A}_{\text{Pamlico}-\text{d},\nu}$ were standardized and transformed back using the means and covariance matrices from Table 2.

Using the $\mathbf{A}_\nu$ sub-matrices, the fixed matrices $\mathbf{H}_k$ and **B**, and a random error three-mode matrix **E**, three-mode data were created using Equation (3). Matrix **E** was created by drawing random values from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 0.463$ (the estimated error variance from the original Blue Crab data) for each entry. To give an impression of the magnitude of this error variance: the proportion of explained variance of the Tucker2 model used here is $R^2 = 0.571$. The categorical grouping variable had values 1, 2, and 3, representing the three subgroups of Crabs. Using the above procedure, 100

replicated complete data sets of $N = 138$ were drawn, henceforth denoted as the *complete data*.

2.1.2. Creating the incomplete data

For each of the 100 complete data sets, different percentages of data points were deleted under different missingness mechanisms.

2.2. Design of the simulation study

2.2.1. Independent variables

2.2.1.1. Missingness mechanism. Missing data were created using three missingness mechanisms: MCAR, MAR, and NMAR. Missing data according to MCAR were created by randomly removing a subsample of data points, with all data points having equal probability. Under MAR, missingness depended on the grouping variable: for diseased crabs from Pamlico River the probability of missing data was three times higher than for crabs from Albemarle, and for healthy crabs from Pamlico River, the probability of missing data was two times higher than for crabs from Albemarle. Given these probability ratios, a number of data points were removed. Finally, under NMAR, for entries in data set $\mathbf{Z}$ above 0 the probability of being missing was three times higher than for entries below zero. Given these probability ratios, a number of data points were randomly removed.

2.2.1.2. Percentage of missingness. 5%, 10%, 20%, and 40% missing data were studied.

2.2.1.3. Missing-data method. Nine missing-data methods were studied (Table 1). Applying PAN to the situation of Crabs × Tissue types × Trace-elements means that tissue types are the model variables and trace elements are nested within crabs; applying PAN-R here means that trace elements are the model variables and tissue types are nested within crabs.

2.2.2. Dependent variables

Van Ginkel and Kiers [40], Van Ginkel and Kroonenberg [29], and Van Ginkel et al. [41] used *Root Mean Squared Bias* (RMSB) of the component loadings as a quality measure of a PCA solution (also, see Bernaards and Sijtsma [42], who used a similar measure). The RMSB can be adapted for the parameters in the Tucker2 model.

In PCA, the problem with comparing sample component loadings with population loadings is that in the sample, the order of components may have changed and that for some components, the signs of the loadings may have been reversed compared to the population component solution. The same may happen in a Tucker2 analysis for the sample estimates of $\mathbf{A}_\nu$ and $\mathbf{B}_\nu$, which we will from now on denote $\hat{\mathbf{A}}_\nu$ and $\hat{\mathbf{B}}_\nu$. To properly align the sample loadings with the population loadings in PCA, the sample solution must be rotated towards the population solution first, using a Procrustes rotation [32,33].

Rotating the estimated core matrix, denoted $\hat{\mathbf{H}}_\nu$, optimally to the population matrix $\mathbf{H}$ is more problematic. Firstly, the Procrustes rotation can only be applied to two-mode matrices while $\hat{\mathbf{H}}_\nu$ is a three-mode matrix. Secondly, when all matrices $\hat{\mathbf{A}}_\nu$, $\hat{\mathbf{B}}_\nu$, and $\hat{\mathbf{H}}_\nu$ are independently rotated, the model that they form defined by Equation (3) gets lost. Instead, the optimally rotated core matrix from a sampled data set ν, denoted, $\hat{\mathbf{H}}^*_\nu$, is constructed

by using the already optimally rotated versions of $\hat{\mathbf{A}}_\nu$ and $\hat{\mathbf{B}}_\nu$, denoted $\hat{\mathbf{A}}_\nu^*$ and $\hat{\mathbf{B}}_\nu^*$, and by using Equation (4).

Once $\hat{\mathbf{A}}_\nu^*$, $\hat{\mathbf{B}}_\nu^*$ and $\hat{\mathbf{H}}_\nu^*$ have been obtained, two fit measures are determined, both based on the root mean squared bias measure. The first fit measure is an RMSB measure for **H** and **B** jointly. Define

$$\mathbf{F} = \mathbf{B}[\mathbf{H}_1, \ldots, \mathbf{H}_k] \tag{6}$$

as a population three-mode matrix which is a generalization of the matrix of the component loadings in two-mode PCA (Equation (2)), and define

$$\hat{\mathbf{F}}_\nu^* = \hat{\mathbf{B}}_\nu^*[\hat{\mathbf{H}}_{1,\nu}^*, \ldots, \hat{\mathbf{H}}_{k,\nu}^*] \tag{7}$$

as the corresponding optimally rotated solution of data set ν. Now suppose f_{jqk} is the entry in **F** for trace element j, component q, and tissue type k, and $\hat{f}_{jqk,\nu}^*$ is the corresponding loading of an incomplete sampled data set ν, under a specific missingness mechanism, percentage of missingness, and using a specific missing-data method. The root mean squared bias of **F** for the νth data set is defined as follows:

$$RMSB(\mathbf{F})_\nu = \sqrt{\sum_{j=1}^{J}\sum_{q=1}^{Q}\sum_{k=1}^{K}(\hat{f}_{jqk,\nu}^* - f_{jqk})^2 / JQK}. \tag{8}$$

The value of RMSB$(\mathbf{F})_\nu$ can be interpreted as the root of the squared distance of the entries in $\hat{\mathbf{F}}_\nu^*$ of the sampled data to the entries of the true matrix F_ν in the multidimensional space, averaged over J trace elements, Q components, and K tissue types.

Likewise, the RMSB of matrix **A** is computed. Suppose $\mathbf{A}_{g\nu}$ is the true matrix of component scores of the crabs in completed data set ν with an entry $a_{ip,g\nu}$ for crab i in group g, and component p, and $\hat{\mathbf{A}}_{g\nu}^*$ is the rotated version of $\hat{\mathbf{A}}_{g\nu}$ in data set ν with an entry $\hat{a}_{pk,g\nu}^*$ for crab i in group g and component p. The root mean squared bias of **A** for group g in replicated data set ν, is defined as follows:

$$RMSB(\mathbf{A})_{g\nu} = \sqrt{\sum_{i=1}^{I}\sum_{p=1}^{P}(\hat{a}_{ip,g\nu}^* - a_{ip,g\nu})^2}. \tag{9}$$

2.2.3. Statistical analyses

To evaluate the quality of the missing-data methods, the results of the two outcome measures were investigated, RMSB(**F**) using a 4 (Percentage of missingness) × 3 (Missingness mechanism) × 9 (Missing-data method), and RMSB(**A**) using a 3 (Group) × 4 (Percentage of missingness) × 3 (Missingness mechanism) × 9 (Missing-data method) ANOVA. Since all percentages of missingness and missingness mechanisms were simulated in the same 100 replicated original data sets and all missing-data methods were applied to these same 100 replications, all factors were within-subjects factors. For the ANOVA with RMSB(**A**) as outcome variable, Group was a within-subjects factor as well because the groups were all in the same data set. Only small ($0.01 \leq$ total $\eta^2 < 0.06$), medium ($0.6 \leq$ total $\eta^2 < 0.14$), and large effects (total $\eta^2 \geq 0.14$) are discussed, following Cohen's [43] guidelines for effect sizes.

Table 4. ANOVA results for root mean squared bias of three-way matrix **F** and matrix **A**.

Dependent variable	Independent variable	$\nu 1$	$\nu 2$	F	Total η^2
RMSB(**F**)	Percentage × Method	21	2079	6946.277	0.198[c]
	Percentage	3	297	11,480.166	0.685[c]
	Method	7	693	3482.038	0.086[b]
RMSB(**A**)	Percentage × Group	6	594	300.969	0.004[a]
	Missingness mechanism × Group	4	396	72.497	0.105[b]
	Percentage	3	297	10,522.415	0.704[c]
	Group	2	198	78.522	0.048[a]

Note: All p-values are less than 0.001. Only effects with a discernable size are shown.
[a]Small effect.
[b]Medium effect.
[c]Large effect.

3. Results

When analysing the results it turned out that method PAN produced substantially larger values of RMSB(**F**) and RMSB(**A**) than the other methods, causing many higher order interactions with discernable effect sizes. Therefore, it was decided to leave method PAN out of the analyses, resulting in a 4 (Percentage of missingness) × 3 (Missingness mechanism) × 8 (Missing-data method) ANOVA for RMSB(**F**), and a 3 (Group) × 4 (Percentage of missingness) × 3 (Missingness mechanism) × 8 (Missing-data method) ANOVA for RMSB(**A**). Table 4 shows the ANOVA results for the effects that meet Cohen's criteria for discernable effect sizes, for both RMSB(**F**) and RMSB(**A**). For RMSB(**F**) the effects that meet these criteria are the interaction of Missing-data method × Percentage of missingness, Missing-data method, and Percentage of missingness. For RMSB(**A**) the interactions of Percentage of missingness × Group and Missingness mechanism × Group, the effects of Group and Percentage of Missingness were discernible. Each of these effects is discussed in more detail below.

3.1. Results of the RMSB of F

3.1.1. Main and interaction effects of method and percentage of missingness

The relevant means and standard errors of RMSB(**F**) are shown in Table 5. For comparison, results of the complete data are shown as well (first row), and the results of method PAN, which was not included in the statistical analysis. As the percentage of missingness increases, RMSB(**F**) increases as well (last row). However, this increase is not the same for all missing-data methods. Especially for the methods that perform the imputations on a dataset in wide format (Reg-W, PMM-W) and method PAN, the increase in RMSB(**F**) is substantially larger than for the other methods.

The last two columns of Table 5 show the means of the main effect of Missing-data method. The methods are ordered with respect to magnitude of the RMSB(**F**). When averaged over all percentages of missingness (last column), PAN-R is the best performing method and PAN is the worst performing method (mainly caused by the results of 10%, 20%, and 40% missingness; for 5% missing-data PAN is close to the other methods). The next best two methods after PAN-R are the methods based on multiple imputation for separate slices (PMM-S, Reg-S; third and fourth rows), followed by methods based on imputation on a data set in tall format (Reg-T, PMM-T). Another noticeable result is that

Table 5. Root mean squared bias of the three-way matrix **F** for all combinations of method and percentage of missingness, aggregated across all missingness mechanisms.

	Percentageof missingness									
	5%		10%		20%		40%		Total	
Method	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Original	795	4	795	4	795	4	795	4	795	4
PAN-R	797	4	801	3	836	4	1107	4	885	3
PMM-S	800	4	806	4	843	4	1123	4	893	3
Reg-S	799	4	806	3	849	4	1178	4	908	3
PMM-T	799	4	806	3	852	4	1141	4	900	3
Reg-T	802	4	814	3	873	3	1216	4	926	3
EM	826	4	864	4	963	4	1274	5	982	4
PMM-W	803	4	815	3	873	3	1523	6	1003	3
Reg-W	802	3	821	3	911	3	1933	7	1117	3
PAN	811	3	873	3	1221	5	3369	33	1569	9
Total[a]	803	4	816	3	875	3	1312	4	952	3

Notes: Entries have been multiplied by 10^3. Totals represent averages over either methods (rows), percentages of missingness (columns), or both (lower right corner).
[a]Original data and PAN not included.

EM (6th row) performs relatively poorly and that there are only three methods having a larger RMSB(**F**) on average.

3.2. Results of the RMSB of A

When leaving out method PAN from the analysis with RMSB(**A**) as the dependent variable, neither a discernible main effect of method nor interactions of method with the other factors were found. The remaining effects with a discernible size (see Table 4) are discussed below.

3.2.1. Interaction of missingness mechanism and group

Table 6 (upper panel) displays the means and standard errors that involve the main effect of Missingness mechanism × Group. For MCAR, RMSB(**A**) is similar across groups (third row). For MAR, RMSB(**A**) is smallest for Albemarle and largest for Pamlico-d (fourth row). This is not surprising because MAR was simulated such that Pamlico-d had the highest probability of missing data and Albemarle had the lowest probability. Since the percentage of missingness influences the stability of parameter estimates and consequently RMSB(**A**), it makes sense that RMSB(**A**) is highest for the group with the most missing data. For NMAR, RMSB(**A**) was highest for group 3 and lowest for group 1 as well, although the difference between groups were smaller than for MAR.

The main effect of Group was of discernable size as well (see Table 4). For Albemarle RMSB(**A**) was smallest and for Pamlico-d RMSB(**A**) is largest (Table 6, last row). However, as the interaction with Missingness mechanism shows, this is mainly caused by missingness mechanisms MAR and NMAR.

3.2.2. Effects of group, percentage of missingness, and interaction

The means and standard errors of RMSB(**A**) for the effect of group may be found in Table 6 (last row). On average, RMSB(**A**) is lowest for Albemarle and highest for Pamlico-d. As the

Table 6. Root mean squared bias and mean bias of matrices $\mathbf{A}_{\text{Albemarle}}$ to $\mathbf{A}_{\text{Pamlico}-\text{d}}$ for all combinations of method and percentage of missingness, aggregated across all imputation methods.

		Group					
		Albemarle		Pamlico-h		Pamlico-d	
Dependent variable		*M*	SE	*M*	SE	*M*	SE
RMSB(**A**)	Original data	2071	13	2071	13	2071	13
	Missingness mechanism						
	MCAR	2390	12	2474	15	2479	13
	MAR	2288	14	2489	15	2632	15
	NMAR	2348	14	2501	14	2561	15
	Percentage of missingness						
	5	2131	13	2160	14	2176	14
	10	2193	13	2243	13	2281	13
	20	2325	13	2442	14	2509	13
	40	2719	14	3106	16	3263	15
	Total[a]	2342	12	2488	13	2557	12

Note: Totals represent averages over missingness mechanism. Entries have been multiplied by 10^5.
[a]Original data and method PAN not included.

percentage of missingness increases, RMSB(**A**) increases as well. These differences become larger as the percentages of missingness increases (rows 7–10). Finally, the main effect of percentages of missingness shows that for 5% RMSB(**A**) is smallest ($M = 2155 \times 10^{-5}$, $SE = 8 \times 10^{-5}$), followed by 10% ($M = 2239 \times 10^{-5}$, $SE = 8 \times 10^{-5}$) and 20% missingness ($M = 2425 \times 10^{-5}$, $SE = 7 \times 10^{-5}$); for 40% missingness RMSB(**A**) is largest ($M = 3029 \times 10^{-5}$, $SE = 9 \times 10^{-5}$).

4. Application to an empirical data set

The above simulations have two disadvantages. Firstly, the data were simulated according to a model. In practice however, data do not behave like a model. Secondly, the performance of the methods is expressed using an overall quality measure (RMSB), but this does not say anything about bias of individual parameters (i.e. individual entries of the matrices **A**, **H**, and **B**). On the other hand it is infeasible to report simulation results of that many parameters. Additional to the simulations (some of) the missing-data methods were applied to an empirical data example. The presentation of the analyses of this empirical data set is a compromise between showing the performance of the missing-data methods at the level of the individual parameters on the one hand, and not reporting too many simulation results on the other hand, with the additional advantage that it also gives an impression of how the missing-data methods behave in real data, rather than simulated data.

The particular data set originates from the Centre of Child and Family Studies of the Department of Education, Leiden University, and was used in earlier studies [44,45]. They will be referred to as the *Strange Situation* data. In this data set, $N = 326$ infants (first mode) are measured on five variables measuring the child's reaction to a strange situation (second mode) during two so-called *reunion episodes* (third mode). The five variables are *Proximity seeking*, *Contact maintaining*, *Resistance*, *Avoidance*, and *Distance Interaction*. Additionally, there is one background variable measuring the attachment style of the child in three types: *Avoidant* (A), *Secure* (B), and *Resistant* (C) [46]. In Ainsworth's categorization a fourth category exists as well, namely *Disorganized/disoriented* (D), but this category did not appear in the data. For more details, we refer to [46].

Table 7. Results of the **B** and **H** matrix of the Tucker2 model applied to the Strange Situation data [44], shown for both the original data and the missing-data methods.

		Component number second way							
		Original data		EM		PMM-S		PAN-R	
Matrix	Variable	1	2	1	2	1	2	1	2
B	Proximity seeking	−522	185	−517	204	−530	167	−524	200
	Contact maintaining	−537	−24	−527	−31	−533	−48	−543	−49
	Resistance	−270	−737	−271	−734	−273	−690	−271	−717
	Avoidance	397	−594	404	−587	395	−642	394	−605
	Distance interaction	457	263	468	272	452	283	450	272
	Component number first way								
$\mathbf{H}_1$	1	25,793	1815	26,265	2782	24,222	2041	24,388	2177
	2	2089	−15,065	1514	−17,434	2019	−14,377	2003	−14,063
	3	−12,265	2072	−13,538	2224	−11,443	1392	−11,371	1518
$\mathbf{H}_2$	1	29,499	1607	30,600	1019	28,519	1548	27,729	1011
	2	−7	−16,386	300	−16,943	53	−14,325	−178	−14,613
	3	10,786	−3472	12,293	−3522	11,367	−1965	9815	−3274
	RMSDO		–	19,437		19,437		19,437	

Note: Entries have been multiplied by 10^3.

Twenty per cent of the scores were randomly removed, using missingness mechanism MAR: for children with attachment style B the probability of missing data on the five variables was twice as high as for children in category A; for children in category C this probability was three times as high. Missing data were handled using the EM algorithm, the multiple-imputation method for two-way data that performed best in the simulation study (PMM-S), and the multilevel imputation method that performed best in the simulation study (PAN-R). The resulting estimated matrices **A**, **H** and **B** were compared with those of the original data. Like in the simulation study, a summary measure of the overall performance was computed for each missing-data method. However, because the population values of **A**, **H**, and **B** are unknown, RMSB cannot be computed. Instead, the *Root Mean Squared Difference with the Original Data* (RMSDO) was used. RMSDO was computed as described in Section 2.2.2, with the only difference that the population parameters f_{jqk} and $a_{ip,g}$ in Equations (8) and (9) were replaced by the estimates of the original data. Also, note that when applied to one data set the replication index ν may be dropped from both equations.

Using deviance and multiway scree plots (Kroonenberg [1, Chapter 8]; Timmerman and Kiers, [39]) it was determined that a 3×2 Tucker2 model had a satisfactory fit for the original data. The same model was fitted on the incomplete data using EM, PMM-S, and PAN-R for handling the missing data. The results of all analyses are given in Tables 7 and 8. The results in both tables show that differences between methods and differences with the original data are small. For matrices **B** and **H** (Table 7) differences in RMSDO between methods are not even visible in the third decimal (last row).

5. Discussion

In this study the influence of several missing-data methods on the results of the Tucker2 model was studied, under different missingness mechanisms and different percentages of

Table 8. Results of the first mode of the Tucker2 model applied to the Strange Situation data [44], shown for both the original data and the missing-data methods, and for each group of attachment style separately.

Matrix	Component number first way	Mean			
		Original	EM	PMM-S	PAN-R
$\mathbf{A}_A$	1	15	14	15	7
	2	15	14	16	7
	3	7	6	6	3
	RMSDO	–	15	12	12
$\mathbf{A}_B$	1	−2	−2	−2	−1
	2	−9	−9	−9	−4
	3	−2	19	−2	−1
	RMSDO	–	25	21	21
$\mathbf{A}_C$	1	−19	−19	−19	−9
	2	20	19	19	9
	3	−4	−4	−4	−2
	RMSDO	–	15	14	14

Note: Entries have been multiplied by 10^3.

missingness. Additionally, results of a Tucker2 analysis on an empirical complete dataset were compared with the results of the same data set, but with some data removed and handled with some of the missing-data methods from the simulation study. As in earlier studies about missing data and PCA [29], the effect of missing data and missing-data methods on the RMSB was small.

For the RMSB of **F**, effects with discernable size were found for Percentage of missingness, Missing-data method, and the interaction of both. As the percentage of missingness increased, RMSB(**F**) increased as well, but the increase was not the same for all missing-data methods. For example, for multilevel imputation with trace elements nested within crabs (data in $IJ \times K$ format), the increase was larger than for multilevel imputation with tissue types nested within crabs (data in $IK \times J$ format).

5.1. Root mean squared bias of F

For RMSB(**F**) it was found that multilevel imputation with trace elements nested within crabs (PAN-R; $IJ \times K$ format) was the best performing method and multilevel imputation with tissue types nested within crabs (PAN; $IK \times J$ format) was the worst performing method. Of the multiple-imputation methods for two-mode data, methods based on predictive mean matching had smaller RMSB(**F**) than their corresponding regression methods. Furthermore, methods based on imputation for separate slices (Reg-S and PMM-S) performed best, followed by methods based on imputation on a tall data set (Reg-T and PMM-T). Methods based on imputation on a wide data set (Reg-W and PMM-W) had the largest RMSB(**F**) of the methods for two-mode data, and were more sensitive to increasing percentages of missingness than the other methods. The relatively poor performance of the latter two methods is probably due to the relatively small sample size and large number of variables, which may increase the risk of overfitting. This overfitting may become even worse for large percentages of missingness than for small percentages of missingness, because the parameters of the imputation model have to be estimated from even less information, but with the same number of variables. Treating different occasions as separate

cases, like methods Reg-T and PMM-T do, may ignore the dependence among occasions, but a larger number of cases with a sufficiently small number of variables is gained in return, which is apparently more important for gaining results with small RMSB than taking into account the dependence among occasions.

Surprisingly, the EM algorithm had a relatively large RMSB(**F**), but this did not seem to be related to missingness mechanism. It was expected that the EM algorithm would perform worse for MAR (and NMAR) than for MCAR because the EM algorithm does not take into account dependence of the missing data on auxiliary variables. However, no relevant interaction was found between Missing-data method and Missingness mechanism.

Another surprising result was that PAN had the largest RMSB(**F**) in general, while this method was supposed to accurately model the hierarchical structure of three-mode data. In a pilot study, it was already found that multilevel imputation using Fully Conditional Specification broke down for large percentages of missing data when tissue types were nested within crabs and trace elements were measured as separate variables (divergence of the imputations to values far beyond the range of the data). This problem was probably caused by the fact that there is little information available for estimating the random effect of tissue type. The same problem may occur in PAN, although to a lesser degree.

Another finding that supports this theory is that method PAN-R, in which tissue types are represented by different variables and trace elements are nested within crabs, had the smallest RMSB(**F**) of all methods. When there is more information about the random effect and there are only few variables, the problem of sparse data may disappear and multilevel imputation will accurately generate imputed values for data with a hierarchical structure.

5.2. Root mean squared bias of A

Leaving out the worst performing method PAN of the analyses, no discernible effect of Missing-data method was found for RMSB(**A**). Apparently, matrix **B** and core array **H** are more influenced by the missing-data method than matrix **A**.

For RMSB(**A**) there was an interaction of group and missingness mechanism. When the data were MAR and NMAR, RMSB(**A**) was smallest for Albemarle and largest for Pamlico-d. Considering the way missingness was simulated, it makes sense that for MAR this pattern was found, because under MAR the simulated Albemarle crabs had the smallest percentage of missingness, and the simulated Pamlico-d crabs had the largest percentage of missingness. More missing data leads to less accuracy of the parameter estimates, resulting in larger values of RMSB(**A**). For NMAR it is less obvious how it results in the same pattern of smallest RMSB(**A**) for Albemarle and largest RMSB(**A**) for Pamlico-d because under NMAR, missingness was related to the value of the missing-data point itself, which is not directly linked to the values of **A**. Furthermore, as the percentage of missingness increased, RMSB(**A**) increased as well.

5.3. Scaling and data format in multilevel imputation

Although method PAN-R generally produced the best results, it should be noted that the simulated data were based on the *standardized* Blue Crab data. In the Blue Crab data, this standardization was necessary because the trace-element levels differed substantially

across different trace elements. Because of this standardization the nesting of trace elements within Crabs like method PAN-R does, caused no problems. However, when variables are not measured on the same scale and have substantially different variances, this may cause problems when variables are nested within cases. The multilevel model assumes a random error term with a constant variance across variables. This assumption is violated when variables with substantially different variances are nested within cases.

To test this conjecture, we also studied method PAN-R when the simulated data were rescaled to the original scales of the Crab data and standardizing the resulting imputed data sets after imputation (results not shown). This option produced substantially larger RMSB(**F**) and RMSB(**A**) than the other methods. This shows that PAN-R will only work when variables have similar means and variances. Thus, our general advice when using multilevel imputation is to use a format of the three-mode data such that the mode with the most levels is nested within cases, and the mode with the least levels is represented by separate variables, regardless of whether the resulting imputation model conceptually makes sense. If the mode with the most levels has substantially different variances across levels, it is advised to standardize the data prior to imputation, and (possibly) transform the data back afterwards.

5.4. Effects of missingness

It was expected that under MAR, missing-data methods that take into account the relevant auxiliary variables (such as multiple imputation) would produce unbiased results and that methods that do not (such as the EM algorithm), would produce biased results. However, neither a relevant main effect of missingness mechanism was found, nor a relevant interaction between missingness mechanism and missing-data method. This implies that the relatively poor performance of the EM algorithm cannot be attributed to violations of MAR for the specific analysis.

To understand these counterintuitive results one must keep in mind that neither RMSB(**F**) nor RMSB(**A**) are measures for bias, but for both bias and inaccuracy. When both are large, either there is systematic under- or overestimation of the parameters, causing large squared deviations of the estimated parameters from the real parameters, or the estimated parameters are unstable, causing large squared deviations from the real parameters as well. Since the performance of the EM algorithm did not depend on missingness mechanism, it may be concluded that the EM algorithm probably produces less stable results than multiple-imputation methods do (i.e. more variability of the estimates in matrices **A**, **B**, and **H**).

It is not clear to us what could have caused this larger instability than that of the multiple-imputation methods. A possible explanation is that the EM algorithm handles the missing data and estimates the Tucker2 model simultaneously. This additional step of handling the missing data in the estimation process may introduce additional instability in the parameter estimates. Multiple imputation on the other hand completes the data multiple times first, so that each Tucker2 model estimated for each of the imputed data sets is estimated with more stability than the solution of the EM algorithm. Additionally, by pooling the results of the M solutions into an overall solution, the variation among the solutions of the M imputed data sets is removed, causing even more stability. However, this explanation is largely speculative and more research is needed.

Instability may also explain the interaction of missingness mechanism and group for RMSB(**A**). Under MAR and inclusion of the relevant auxiliary variable(s) in the missing-data method, results are expected to be unbiased. In our study we found that under MAR, groups with a large percentage of missingness (Pamlico-d) had larger RMSB(**A**) on average than groups with a low percentage of missingness (Albemarle). However, this difference in RMSB(**A**) was probably caused by increased inaccuracy and not by bias.

To summarize, if we would like to draw conclusions about possible bias of the EM algorithm under MAR, and lack of bias under MAR for multiple-imputation methods, we would have to use quality measures not based on squared deviations but on non-squared deviations. Besides root mean squared bias, Van Ginkel and Kroonenberg [29], Van Ginkel et al. [41], and Van Ginkel and Kiers [40] studied another quality measure for the component loadings, namely *mean bias*. This measure was simply the deviation of the estimated component loading from the corresponding population loading, averaged across all variables and components. In the current study, mean bias was studied but not discussed because unlike in the other studies mentioned, the mean bias gave results that were not clearly interpretable, probably because it was heavily influenced by positive and negative biases cancelling each other out when averaged over all variables (i.e. trace elements).

5.5. Empirical data set

To get an impression of the extent to which the individual entries in matrices **A**, **H**, and **B** were influenced by the missing-data methods, the EM algorithm, the best performing multiple-imputation method for two-mode data (PMM-S) and the best performing multilevel imputation method (PAN-R) were applied to the Strange Situation data [44,45] with simulated missingness. Results obtained using the several missing-data methods differed little from the results of the same data set without missing data, and differed little from each other as well.

5.6. Generalizability to other three-mode models and other data properties

In this paper only the results of the Tucker2 model were studied. However, we expect that the results of the Tucker3 model will not differ much from the results in this study because the Tucker3 model is very similar to the Tucker2 model. The most important difference between the Tucker2 and the Tucker3 model is that the Tucker3 model has a component matrix for the third mode as well, denoted **C** [1, p. 54–57]. Although no rules for combining the **C** matrices from several imputed datasets have been defined, the GPA approach described in Section 1.4.2 of Kroonenberg and Van Ginkel [12] may be applied to **C** as well, and the core matrix may then be constructed using a similar procedure as in Equation (5). As the Parafac model has different characteristics, such as uniqueness and possibilities of degeneracy, this procedure may not automatically be applicable to Parafac models and therefore Parafac models deserve a separate study.

Furthermore, the model parameters and other properties of the simulated data sets (such as sample size and number of variables) were based on one particular data set (i.e. the Blue Crab data). The question is to what extent the results can be generalized to data sets with other properties. Although the performance of the missing-data methods was illustrated using an empirical data set with different properties as well (Strange

Situation data), the properties of this data set were not part of a simulation study so it remains unclear to what extent the results generalize to other data sets from the same population.

For two-mode (principal component) analysis, there is some evidence that model parameters, sample size, and number of variables have little influence on the results. Van Ginkel et al. [41] studied the influence of various missing-data methods on the results of PCA using the properties of three different existing data sets for their simulation models. The effect of these different properties on the results was negligible. Although we cannot be certain that these findings generalize to three-mode analysis as well, we have no reason to assume that the choice for sample size and the number of variables would heavily influence the findings in this study.

The choice of the model parameters on the other hand, could be of influence on the results, especially when this concerns the strength of association between the slices of the third mode. It is expected that the performance of methods not taking into account dependence among slices (multiple-imputation methods for two-mode data applied to separate slices, or applied to a data set in long format) will become worse as the dependence among slices becomes stronger. Therefore, we must be cautious generalizing the findings of these methods to other three-mode data sets. However, these methods were not the best performing methods to start with.

Another aspect of the Crab data that may have limited generalizability to other three-mode datasets is the number of levels of the third mode (K). The Crab data are relatively exceptional for a three-mode dataset in the sense that $K < IJ$. The question is to what extent similar results may be found when K is larger. However, the choice for the Crab data was based on the availability of a background grouping variable (Albemarle Sound – healthy, Pamlico River – healthy, and Pamlico River – diseased), which could perfectly serve the purpose of simulating MAR.

5.7. Future research

There are some other aspects about missing data in three-mode analysis that were not studied but which may require further studying. Examples are attrition (some cases having complete occasions missing), deviations from normality, and the Parafac model as the statistical analysis of interest. These topics could be studied in future research.

6. Conclusion

To summarize, multiple imputation seems to be a good alternative for the EM algorithm in the Tucker2 model, especially multilevel imputation with the mode having the largest number of levels nested within cases, and the mode with the least levels represented as separate variables, provided that the data are standardized first (PAN-R). Whenever multiple imputation under the multilevel model is either too complicated for, or not available to the researcher, one could use multiple-imputation methods applied to data set in long format, or applied to separate slices, but avoid using multilevel imputation with the mode having the least levels nested within cases (PAN). Finally, more research on the EM algorithm is needed to find out more about the possible causes of its seemingly less stable results than those of the multiple-imputation methods.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Kroonenberg PM. Applied multiway data analysis. Hoboken, NJ: Wiley; 2008.
[2] Kroonenberg PM, De Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika. 1980;45:69–97.
[3] Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika. 1966;31:279–311.
[4] Little RJA, Rubin, DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.
[5] Rubin DB. Inference and missing data. Biometrika. 1976;63:581–592.
[6] Dempster AP, Laird NM, Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B. 1977;39:1–38.
[7] Weesie J, Van Houwelingen H. GEPCAM user's manual: generalized principal components analysis with missing values. Utrecht, the Netherlands: Institute of Mathematical Statistics, University of Utrecht; 1983. (Technical report.)
[8] Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
[9] Tucker LR. Relations between multidimensional scaling and three-mode factor analysis. Psychometrika. 1972;37:3–27.
[10] Harshman RA. Foundation of the PARAFAC procedure: models and condition for an "explanatory" multi-modal factor analysis. UCLA Work Pap Phon. 1970;16:1–84.
[11] Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika. 1970;35:283–319.
[12] Kroonenberg PM, Van Ginkel JR. Combination rules for multiple imputation in three-way analysis illustrated with chromatography data. Curr Anal Chem. 2012;8:224–235.
[13] Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC Press; 2012.
[14] Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
[15] Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Amer Statist Assoc. 1987;82:528–540.
[16] Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul. 2006;76:1049–1064.
[17] Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681–694.
[18] Rubin, DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat. 1986;4:87–94.
[19] Marshall A, Altman DG, Holder, RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112. DOI:10.1186/1471-2288-10-112
[20] Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data with prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7.
[21] SPSS Inc. IBM SPSS Statistics 23.0 for Windows. Chicago, IL: SPSS Inc.; 2015.
[22] Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Soft. 2011;45:1–67.
[23] Yuan Y. Multiple imputation using SAS software. J Stat Soft. 2011;45:1–25.
[24] Goldstein H. Multilevel models in educational and social research. London: Charles Griffin & Company Ltd; 1987.
[25] Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed models with missing values. J Comput Graph Stat. 2002;11:437–457.
[26] Van Buuren S. Multiple imputation of multilevel data. In: Hox JJ, Roberts JK, editors. The handbook of advanced multilevel analysis. Milton Park: Routledge; 2011. p. 173–196.

[27] Schafer JL. Pan: Multiple Imputation for Multivariate Panel or Clustered Data. R package version 0.3. http://CRAN.R-project.org/package = pan; 2011.
[28] Van Ginkel JR, Kroonenberg PM. Analysis of variance of multiply imputed data. Multivariate Behav Res. 2014;49:78–91.
[29] Van Ginkel JR, Kroonenberg PM. Using generalized procrustes analysis for multiple imputation in principal component analysis. J Classif. 2014;31:242–269.
[30] Gower JC. Generalized procrustes analysis. Psychometrika. 1975;40:33–51.
[31] Ten Berge JMF. Orthogonal procrustes rotation for two or more matrices. Psychometrika. 1977;42:267–276.
[32] Green BF. An orthogonal approximation of an oblique structure in factor analysis. Psychometrika. 1952;17:429–440.
[33] Gower JC. Statistical methods of comparing different multivariate analyses of the same data. In: Hodson FR, Kendall DG, Tautu P, editors. Mathematics in the archaeological & historical sciences. Edinburgh: Edinburgh University Press; 1971. p. 138–149.
[34] Hubert M, Van Kerckhoven J, Verdonck T. Robust PARAFAC for incomplete data. J Chemometr. 2012;26:290–298.
[35] Louwerse DJ, Smilde AK, Kiers HAL. Cross-validation of multiway component models. J Chemometr. 1999;13:491–510.
[36] Tian T, McLachlan GJ, Dieters MJ, Basford KE. Application of multiple imputation for missing values in three-way three-mode multi-environment trial data. PLOS One. 2015;10. DOI:10.1371/journal.pone.0144370.
[37] Gemperline PJ, Miller KH, West TL, Weinstein JE, Hamilton JC, Bray JT. Principal component analysis, trace elements, and blue crab shell disease. Anal Chem. 1992;64:523A–532A.
[38] Kroonenberg PM, Basford KE, Gemperline PJ. Grouping three-mode data with mixture methods: the case of the diseased blue crabs. J Chemom. 2004;18:508–518.
[39] Timmerman ME, Kiers, HAL. Three-mode principal components analysis: choosing the numbers of components and sensitivity to local optima. Br J Math Stat Psychol. 2000;53:1–16.
[40] Van Ginkel JR, Kiers HAL. Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: a multiple-imputation approach. Br J Math Stat Psychol. 2011;64:498–515.
[41] Van Ginkel JR, Kroonenberg PM, Kiers HAL. Missing data in principal component analysis of questionnaire data: a comparison of methods. J Stat Comput Simul. 2014;84:2298–2315.
[42] Bernaards CA, Sijtsma K. Factor analysis of multidimensional polytomous items response data suffering from ignorable item nonresponse. Multivariate Behav Res. 1999;34:277–313.
[43] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
[44] Kroonenberg PM, Basford KE, Van Dam M. Classifying infants in the strange situation with three-way mixture method clustering. B J Psychol. 1995;86:397–418.
[45] Van IJzendoorn MH, Goossens FA, Kroonenberg PM, Tavecchio, LWC. Dependent attachment: B4-children in the strange situation. Psychol Rep. 1985;57:439–451.
[46] Ainsworth MDS, Blehar MC, Waters E, Wall S. Patterns of attachment. A psychological study of the strange situation. Hillsdale, NJ: Erlbaum; 1978.