RESTRICTIVE IMPUTATION OF INCOMPLETE SURVEY DATA

GERKO VINK

Gerko Vink
Utrecht University, Utrecht

Restrictive Imputation of Incomplete Survey Data

RESTRICTIVE IMPUTATION OF INCOMPLETE SURVEY DATA

IMPUTATIE VAN INCOMPLETE STEEKPROEFDATA ONDER RESTRICTIES
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 13 maart 2015 des middags te 4.15 uur

door

Gerrit Vink

geboren op 17 juli 1984
te Apeldoorn

Promotor: Prof.dr. S. van Buuren
Copromotor: Dr. J. Pannekoek

Beoordelingscommissie:

Prof.dr. P.G.M. van der Heijden
Prof.dr. G. Molenberghs
Prof.dr. A.W. Taris
Prof.dr. J.W.R. Twisk
Prof.dr. A.G. de Waal

Vink, Gerko
Restrictive Imputation of Incomplete Survey Data
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 978-90-393-6300-3
Druk: GVO drukkers & vormgevers B.V. | Ponsen & Looijen
Cover design by Gerko Vink

To everyone without whom this dissertation
would have been completed earlier

Acknowledgements

As with every finished project, proper thanks are due. I am very grateful for the support of my supervisors. Stef, your mentoring has been critical to my scientific growth and our working together has given me unique opportunities. Thank you for challenging me in times I needed it. Jeroen, without your guidance and vast knowledge of official statistics, this project would have gotten elsewhere. You are the reason I felt at home away from home at Statistics Netherlands. Between you both, I dare to say to have had the best supervision possible.

Laurence, I thank you for your daily supervision in the first years of this project. You have taught me that any message is completely redundant if not delivered properly.

I would like to thank the people at Statistics Netherlands and everyone else with whom I have collaborated over the years. Even though this thesis may not show any direct results of our collaboration, I can assure you that you have all had an influence. In particular I would like to thank Goran. We have proven that long distance collaborations are perfectly executable over a Skype connection.

I thank everyone from M&S at Utrecht University. You have all to some extend made me feel at home much in the same way a village is connected to its idiot. I appreciated all the usual stuff: the drinks, the dinners, the pancakes, the coffee. But most of all, I enjoy the genuine interest in one-another, the (long) conversations, the pranks (no regrets), and the fact that people would gladly work past five o'clock.

I am grateful for the people that have had to share office space with me. Thomas, Maria, Shahab, Suzanne, and Nino. You are the best roommates one could wish for. Charlotte, Els, Flip, Irene, Jesper, Joran, Kevin, Marieke, Maryam, Nijs, Noémi, Peter, Rebecca, and Rens: you are (or have been) the best extended roommates!

Kevin, I consider us to be partners in crime. Thanks for your help, humor and friendship in the past years. I literally look up to you.

Shahab, you are my scientific big brother. Thank you for keeping me grounded. I could not wish for a better friend.

My friends and family are important to me. The good old ones, the newer ones and the ones acquired through relational mergers: thank you all for putting up with me in the past years. Also, thank you for making me feel funnier than I really am.

My parents I thank for giving me the opportunity to develop a scientific mind. You could not have given me a greater gift.

Finally, I thank my wife Aly. They say that love conquers all. They are right: From the moment I met you, our life has been smooth sailing. To camel rides and wearing kilts!

Utrecht, January 2015
Gerko Vink

Contents

1

Introduction

Missing data form a ubiquitous source of problems that most scientists or researchers cannot escape. For example, in survey applications, such as in social sciences or in official statistics, where vast amounts of data are collected, respondents often neglect to answer one or more items. Although the field of missing data has been largely developed in the past decades with survey applications in mind - Rubin's (1987) initiating book is not without purpose named 'Multiple Imputation of Nonresponse in Surveys' - missingness occurs throughout the whole of science.

For example, in astrophysics, properties of hardly observable, distant objects are often not directly observable. However, no matter their distance, such objects always have neighboring objects around them and take their part in a cluster or solar system within their galaxy. Examining such constellations of objects in space as a whole, enables astrophysicists to deduce information about the objects of interest, or at least quantify their properties with a certain level of confidence.

The procedure astrophysicists follow is analogous to the approach that can be taken with missing data in survey research. Most of the times information about respondents (objects) can be inferred about the people around them (solar system) or, if needed, about the information that is observed in the data as a whole (galaxy). Drawing inference based on data with missing values implies that at least some information is observed, which conveniently redefines the problem of missingness into a problem of incompleteness. After all, incomplete data sounds much more like a solvable puzzle than missing data.

1.1 Drawing inference on incomplete data

Let us assume the classical research scenario where a researcher operationalizes a problem, formulates a research question, collects data, analyzes the data and finds an answer to the research question. Any occurrence of missing values in the acquired dataset has serious consequences for the analysis phase, as most analyses require the data to be completely observed. Suppose that the researcher is well aware that the

missingness in the dataset should never be ignored - perhaps only because ignoring the missing values will yield lower statistical power. This leaves the researcher at a junction of two very different paths in order to obtain an answer to the research question.

The first path leads to the answer directly. In general we could say that estimation procedures that follow the first path, such as maximum likelihood, weighting and full Bayesian estimation techniques, aim to identify the population parameters that are most consistent with the observed data. These estimation approaches adhere to some sort of model, which may be implicit in the case of weighting, or explicit in the case of maximum likelihood or Bayesian estimation. To allow for the missing data during analysis, the estimation methods must be adapted to deal with incomplete data and that missing values need to be integrated out of the analysis model. As a consequence, some of the cases in the data will contribute more information to answering the research question than others.

The second path, imputation, first solves the missing data problem. With imputation, some estimation procedure is used to impute (fill in) each missing datum, resulting in a completed dataset that can be analyzed as if the data were completely observed.

When only one value is imputed (single imputation), uncertainty about the imputations is not reflected in the imputed data set and specific methods for variance estimation that take imputed values into account need to be employed. As a more versatile way to solve this, uncertainty about the imputed values can be taken into account by performing multiple imputation (MI). With MI, each missing datum is imputed $m \geq 2$ times, resulting in m completed datasets. At least 2 imputations are warranted to reflect the uncertainty about the imputations, although performing more imputations is often advisable. The m datasets are then analyzed by standard procedures and the analyses are combined into a single inference.

In this dissertation, only MI is considered. The choice for MI is based on the following arguments. First, MI is quickly becoming more popular and is easily one of the most utilized methods for dealing with nonresponse in many domains of statistics. This can also be seen in the growing number of books and conferences that consider MI. A possible explanation for MI's popularity has to do with separating the missing data problem from the analysis stage. As a result, inference using MI is relatively straightforward to obtain and easy to comprehend, properties that may be particularly appealing to applied researchers.

Second, one can imagine that obtaining inference with direct estimation procedures becomes increasingly more complicated when modeling the data becomes more challenging, such as with large amounts of variables or complex univariate or multivariate distributions. With multiple imputation, such complexities are mostly applicable to the imputation stage, making the analysis stage relatively straightforward. In other words, once satisfactory imputations are obtained it is not too difficult to answer the research question.

1.2 Obtaining plausible imputations

In this dissertation only plausible imputations - imputations that could be real values if they had been observed - are considered to be satisfactory. This definition of plausibility considers the position of imputed values, given the data. For ordinary datasets, this means that plausibility must be considered in two directions: the incomplete variable (column) and the remainder of the measurements of the respondent (row). In other words, plausibility should consider the imputed value and the relation that imputed value has to other (observed and imputed) values in the data. For example, if variables sum up to a certain total, only those imputations are plausible that obey the structure of the sum.

For continuous data, the normal linear regression imputation model is a very basal approach to obtaining multiply imputed values. With normal linear regression imputation, imputations are drawn (with error) from a regression model, such that incomplete outcomes are predicted based on observed (or imputed) values in a set of predictors (see e.g. (Rubin, 1987, p. 167)). Assumptions about the type of data and the shape of the distributions in the data are explicitly made and deviations from these assumptions may yield invalid inference. In practice, other types of data are often encountered and the normal imputation model may not yield plausible imputations.

When performing multiple imputation, there are properties of the data that one would like to preserve during imputation. Sometimes these properties are not directly visible (e.g. intricate multivariate relations) and sometimes these properties are very explicit (such as summations or polynomials). Preserving such data properties limits the space where imputations can be sampled from. As a result, restrictions are put on the model that can be used for obtaining plausible imputations.

1.3 Aim

This dissertation focuses on finding plausible imputations when there is some restriction posed on the imputation model. In these restrictive situations, current imputation methodology does not lead to satisfactory imputations. The restrictions, and the resulting missing data problems are real-life situations that are frequently encountered across different domains of statistics, such as official statistics, social sciences, geology and medicinal sciences. More specifically, imputation strategies that yield plausible imputations are considered for the following restrictive problems.

First, in official statistics highly skewed semicontinuous (or zero-inflated) data are frequently encountered. When imputing these data, the non-negative mixture of continuous values and the point mass (often at zero) need to be considered in such a way that imputations fall within the plausible range of values. Current imputation approaches use multi-step approaches that depend on data-transformations to conform the incomplete data to the imputation model. A single-step imputation solution that does not require data transformations and leads to valid inference and plausible imputations is discussed in Chapter 2.

Second, in many domains in statistics, multilevel (or clustered) data are often encountered. With multilevel data, groups of respondents share common characteristics and can be clustered into classes. This class structure, often summarized in the intraclass correlation coefficient, needs to be taken into account when imputing such data. An imputation approach that provides a straightforward solution for obtaining plausible imputations while taking the multilevel structure of the data into account is discussed in Chapter 3.

Third, applied researchers frequently use squared terms in their analysis models. It is known that the imputation model should embrace all relations of scientific interest. When generating plausible imputations, the relation between the original variable and its squared counterpart needs to be preserved. After all, a squared value that has no relation to its square root, can never be deemed plausible. Chapter 4 proposes an imputation technique for obtaining plausible imputations when the imputation model contains squared terms.

Fourth, in many domains in statistics, compositional data structures are encountered. Compositional data can be defined as a set of parts that obey a certain edit restriction, such that the parts have to sum up to a certain total. Imputing compositional data is challenging because imputations must obey the restrictions in the data while remaining strictly non-negative. Chapter 5 proposes an imputation approach that can handle intricately nested compositional data and provides plausible imputations that adhere to the compositional structure.

Finally, when evaluating imputation approaches, simulations studies are often used. Data are usually sampled from some sort of theoretical distribution that serves as the population. If this is not possible, design-based simulation studies are performed, where data is usually sampled from some 'true' dataset of sufficient size. Both simulation approaches introduce sampling variance, which is not of specific interest when evaluating imputations. Chapter 6 demonstrates a simplification of the conventional pooling rules for multiple imputation in situations where sampling variance is not of interest. These pooling rules are also applicable in situations where the size of the population is restricted and essentially all units in the population have been observed.

1.4 Imputation strategies for multivariate data

Multiple imputation for multivariate data comes in two main flavors: joint modeling (JM) and fully conditional specification (FCS). With JM, imputations are drawn from an assumed joint multivariate distribution. Often a multivariate normal model is used for both continuous and categorical data, although other joint models have been proposed (see e.g. Olkin and Tate, 1961; Van Buuren and van Rijckevorsel, 1992; Schafer, 1997; Van Ginkel et al., 2007; Goldstein et al., 2009; Chen et al., 2011). Joint modeling imputations generated under the normal model are usually robust to misspecification of the imputation model (Schafer, 1997; Demirtas et al., 2008), although transformation towards normality is generally beneficial.

Contrary to JM, multiple imputation by means of FCS does not start from an explicit multivariate model. With FCS, multivariate missing data is imputed by univariately specifying an imputation model for each incomplete variable, conditional on a set of other (possibly incomplete) variables. The multivariate distribution for the data is thereby implicitly specified through the univariate conditional densities and imputations are obtained by iterating over the conditionally specified imputation models.

The general idea of using conditionally specified models to deal with missing data has been discussed and applied by many authors (see e.g. Kennickell, 1991; Raghunathan and Siscovick, 1996; Oudshoorn et al., 1999; Brand, 1999; Van Buuren et al., 1999; Van Buuren and Oudshoorn, 2000; Raghunathan et al., 2001; Faris et al., 2002; Van Buuren et al., 2006). Comparisons between JM and FCS have been made that indicate that FCS is a useful and flexible alternative to JM when the joint distribution of the data is not easily specified (Van Buuren, 2007) and that similar results may be expected from both imputation approaches (Lee and Carlin, 2010).

In this dissertation, new methodology based on FCS is introduced, although comparisons are occasionally made to imputation approaches that utilize some form of joint modeling. The choice for FCS is based on applicability, by avoiding the complex specification and estimation of multivariate models that observe different kinds of restrictions. Because the multidimensional imputation problem is split in multiple unidimensional imputation problems, it is relatively simple to specify imputation models that do not conform to standard multivariate distributions. Moreover, this flexibility in specifying univariate imputation models makes it much easier to adapt imputation models to accommodate for some form of restriction. As a result, the incomplete data can be more efficiently addressed and unique data features can be preserved. For example, in official statistics many restrictions are posed on survey or register data, such as bounds (no unrealistic human age), strict non-negativity (no negative incomes) and conditional restrictions (girls under twelve years of age are not allowed to have children, nor can they be married).

1.5 Current modeling practice

A straightforward implementation of FCS can be found in the MICE algorithm proposed by Van Buuren and Groothuis-Oudshoorn (2000, 2011). The MICE algorithm is a Markov Chain Monte Carlo (MCMC) method, which becomes a Gibbs sampler in situations where the conditional densities are said to be compatible. Compatibility is reached when the joint multivariate distribution has the separate conditional distributions as its conditional densities. For the MICE algorithm, the joint distribution is only implicitly known and compatibility may be difficult to prove. In some situations, compatibility may not actually exists. However, in practice FCS seems to be robust when compatibility conditions are not met (Van Buuren et al., 2006). Recently, Bartlett et al. (2014) introduced a substantive model compatible FCS (SMC-FCS) that ensures that each covariate is imputed from a model which is compatible

with the substantive model. This may be particularly of interest when the substantive analysis model contains non-linearities or interactions.

The MICE algorithm starts with randomly drawing imputations from the observed data. Subsequently, the variables are imputed in a variable-by-variable approach. A single iteration of the algorithm cycles through all incomplete variables.

The number of iterations for the MICE algorithm has to be carefully chosen. In most situations, a low number of iterations appears to be enough (Brand, 1999; Van Buuren et al., 1999), but slow convergence can occur if, for example, the amount of missing data is large or if there is high autocorrelation in the imputation chains. After imputation, convergence of the m multiple imputation chains should be investigated.

The number of imputations is also of importance when doing multiple imputation. Usually, the default amount of imputations in software is set to be as low as three to five. Many authors have investigated the role of m with regard to several criteria, such as the confidence interval, statistical power and the proportion of missingness attributable to the nonresponse (see e.g. Royston, 2004; Graham et al., 2007; Bodner, 2008; White et al., 2011). The work by these authors suggests that it may often be beneficial to set the amount of imputations much larger, although it comes at a cost in terms of data storage and computational time.

In general it holds that using a higher m is always better. This does not necessarily mean that outcomes from resulting analyses will be better. In fact, Schafer (1997) suggests that resources can often be better spent and Schafer and Olsen (1998) indicate that in most situations there is only little advantage to analyzing more than a few imputed datasets. To save computation time and resources, Van Buuren (2012) suggests to set $m = 5$ during model building and to increase m only for the 'actual' imputation stage. However, with computers becoming increasingly faster and data storage solutions becoming more accommodative of large datasets, one can imagine that today's drawbacks in performing more imputations are becoming increasingly less important in the future.

1.6 Choosing an imputation model

The FCS framework allows variables to be imputed under an appropriate model, given the data. For example, continuous variables can be imputed by a Bayesian normal linear imputation model, dichotomous variables by a logistic model and unordered categorical variables can be imputed by a polytomous regression model. In all three applications, imputations are drawn under a formulated model and, like with all modeling efforts with missing data, imputations may be susceptible to misspecification of that imputation model. Although many misspecification problems can be minimized by data transformation beforehand (and backtransforming the imputed data afterwards), it would be preferred if one imputation model could allow for multiple types of data.

This dissertation focuses mainly on imputations generated by means of predictive mean matching (PMM). With PMM, bayesian normal linear regression is used to

obtain predicted values for the observed and missing values in a variable, but the predicted values are not used as imputations directly. Instead, observed and unobserved predictions are 'matched' - usually by minimizing a difference on the predicted values - and the corresponding observed value is chosen as the imputation. This strategy can often be effectively applied to data of all measurement levels, while allowing for imputations that will always fall within the range of plausible observed values (no negative age or income, no unrealistic age). Matching the predicted values and selecting the corresponding observed value as an imputation makes it also an interesting candidate for compound data distributions.

Being able to impute different types of variables by a single imputation strategy is very convenient in practice and should substantially decrease the modeling effort associated with FCS. Also, when sampling from the observed values, there is no need for data transformations to accommodate the imputation model, nor is there a need for postprocessing the imputed values to conform to natural restrictions in the data. This allows for a faster imputation process, a property that is particularly appealing to fields where timely publication of results is considered crucial, such as in official statistics.

Drawing imputed values from the set of observed values should always be done with caution. In situations where the range of observed values in the incomplete sample differs from the range of possible values in the population, imputations generated by strategies like PMM may yield biased inference. Strict modeling approaches, such as bayesian linear regression imputation, may still lead to plausible inference in such situations. Therefore, imputed data should never be taken for granted and imputations should always be critically evaluated.

1.7 Evaluating imputations

Evaluating imputations is of crucial importance and imputed values should never be considered without close examination. For instance, when using an algorithm to generate imputations, the convergence of that algorithm should always be evaluated. A straightforward approach to monitoring convergence is plotting some statistic of interest (usually the mean and standard deviation of the imputed data are used) for each of the multiple imputation chains. The pattern over the iterations should be free of trend, or, in the case of slow convergence, should have reached a stable plateau.

Besides checking convergence of the algorithm, the imputation model and the resulting imputed data should be evaluated too. Most often imputations are generated by means of models that are fitted to the observed data. The fit of these models can be assessed by standard model evaluation tools, such as Q-Q plots and information criteria. Such model evaluation tools generally focus on the fit between the data and the model. However, evaluating the distributional discrepancy between observed and imputed data can often be more informative (Van Buuren, 2012).

Studying the discrepancy between observed and imputed data is a valuable tool to assess plausibility of imputations. Plausible imputations are those that conform to

the distribution of the observed data, not in a sense that the distributions are equal, but rather in the sense that imputations could have been real values had they been observed (Van Buuren, 2012). To evaluate plausibility of imputations, (conditional) distributions can be compared and means, variances, scales and relations between variables can be evaluated. Finally, Abayomi et al. (2008) raise a valid point when they state that researchers should always use a standard of reasonability. Whether observed and imputed values differ from one another or not, the (lack of) discrepancies should always make sense in the context of the problem being studied.

1.8 Outline of the dissertation

This dissertation focuses on plausible value imputation for situations where some restriction is posed on the data. A distinction can be made between three parts. Part I considers univariate strategies for multiply imputing multivariate data by means of FCS. In some situations it can be more convenient to impute two variables at the same time to preserve relations in the data that can otherwise not be preserved. Part II considers such bivariate imputation strategies to deal with missing values. Part III does not consider the imputation process itself, but rather focuses on pooling multiple imputations to obtain a single inference in situations where the size of the population is restricted.

Part I

Univariate imputation

2

Predictive Mean Matching Imputation of Semicontinuous Variables

Summary. Multiple imputation methods properly account for the uncertainty of missing data. One of those methods for creating multiple imputations is predictive mean matching (PMM), a general purpose method. Little is known about the performance of PMM in imputing non-normal semicontinuous data (skewed data with a point mass at a certain value and otherwise continuously distributed).We investigate the performance of PMM as well as dedicated methods for imputing semicontinuous data by performing simulation studies under univariate and multivariate missingness mechanisms. We also investigate the performance on real-life datasets.We conclude that PMM performance is at least as good as the investigated dedicated methods for imputing semicontinuous data and, in contrast to other methods, is the only method that yields plausible imputations and preserves the original data distributions.

2.1 Introduction

Semicontinuous variables consist of a (usually fairly large) proportion of responses with point masses that are fixed at some value and a continuous distribution among the remaining responses. Variables of this type are often collected in economic applications, but can also be found in medical applications. Examples of semicontinuous variables with point masses at zero are income from employment, number of employees or bacterial counts. Semicontinuous variables differ from censored and truncated variables in that the data represented by the zeros, are bona fide and valid, as opposed to the data being proxies for negative values or missing responses (Schafer and Olsen, 1999)

This chapter is published as Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61-90.

2.1.1 Imputation methods for semicontinuous data

In the past decades, the field of imputation has made a major advance. Many model-based imputation procedures have been developed for multivariate continuous and categorical data (Little and Rubin, 2002; Rubin, 1987; Schafer, 1997). Univariate models for modeling semicontinuous data have been developed as well as the Tobit model (Amemiya, 1984; Tobin, 1958) and selection models (Heckman, 1974, 1976). The two-part model seems to be particularly interesting for modeling semicontinuous data. This model presents the data as a two-part mixture of a normal distribution and a point mass (Schafer and Olsen, 1999; Olsen and Schafer, 2001), thereby decomposing the semicontinuous observations into two variables that can be modeled in succession. The two-part model can benefit from transforming the continuous part of the data to normality (White et al., 2011).

Javaras and Van Dyk (2003) introduced the blocked general location model (BGLoM), designed for imputing semicontinuous variables. The BGLoM incorporates a two-part model in the general location model. Expectation-maximization and data augmentation algorithms for generating imputations under the BGLoM have been introduced by Javaras and Van Dyk (2003).

The methods described earlier are based on the multivariate normal distribution. The normal distribution, however, may not accurately describe the data, potentially leading to unsatisfactory solutions (Van Buuren, 2012), which stresses the need for a method without distributional assumptions.

Nonparametric techniques, such as hot-deck methods, form an alternative class of methods to create imputations. In hot-deck methods, the missing data are imputed by finding a similar but observed record in the same dataset, whose observed data serve as a donor for the record with the missing value. Similarity can be expressed, for example, through the nearest-neighbor principle, which aims to find the best match for a certain record's missing value, based on other values in that same record.

A well-known and widely used method for generating hot-deck imputations is *predictive mean matching* (PMM)(Little, 1988), which imputes missing values by means of the nearest-neighbor donor with distance based on the expected values of the missing variables conditional on the observed covariates.

Yu et al. (2007) investigated general purpose imputation software packages for multiply imputing semicontinuous data. Among the software investigated were routines and packages for `SAS [PROC MI, PROC MIANALYZE, IVEware` (Raghunathan et al., 2002)`]`, `R [mice` (Van Buuren and Groothuis-Oudshoorn, 2011) and `aregImpute]` and `Stata [ice` (Royston, 2005)`]`. They concluded that procedures involving PMM performed similar to each other and better than the procedures that assumed normal distributions. PMM not only yielded acceptable estimates, but also managed to maintain underlying distributions of the data (Heeringa et al., 2002; Yu et al., 2007).

Although the research by Yu et al. (2007) is useful, it yields only limited insight in the reasons why PMM works for semicontinuous data. Yu et al. (2007) focus on readily available software implementations, setting aside methods specifically designed for semicontinuously distributed data (Javaras and Van Dyk, 2003; Schafer and Olsen,

1999; Olsen and Schafer, 2001). Even the procedures implementing PMM had different performances, indicating that a distinction must be made between methods and software implementations.

The list of software, as described by Yu et al. (2007) is outdated. New algorithms and packages with support for semicontinuous data have emerged, such as the R-packages `mi` (Su et al., 2011) and `VIM` (Templ et al., 2011). Both methods use an approach to semicontinuous data that is based on the two-part model. `mi`, for example, uses a two-part model where the continuous part is imputed based on log-transformed data. The iterative robust model-based imputation (`irmi`) algorithm from the package `VIM` mimics the functionality of IVEware (Raghunathan et al., 2002), but claims several improvements with respect to the robustness of the imputed values and the stability of the initialized values (Templ et al., 2011).

2.1.2 Goals of this research

Little is known about the practical applicability of PMM on semicontinuous data, and how the method compares to techniques that are specifically designed to handle these types of data. Certain characteristics, such as sample size, skewness, the percentage of zeros and the number of predictors, as well as the strength of relations in the data may play a vital role in the performance of PMM.

We investigate how PMM compares to dedicated methods for imputing semicontinuous data. We thereby concentrate on a comparison between PMM, the two-part model, the BGLoM and the algorithms `mi` and `irmi`. More in particular we investigate how performance is affected by skewness, sample size, the amount of zeros, the percentage missingness and the relations in the data. We also look into the effect of the missing data mechanism on imputation methods for imputing semicontinuous data. We investigate the aforementioned methods in the presence of univariate and multivariate missingness. And, finally, we wonder: is PMM at least as good as a dedicated method when imputing semicontinuous data?

2.2 Imputation methods

2.2.1 Notation and preliminaries

Let $Y = (Y_{obs}, Y_{mis})$ be an incomplete semicontinuous variable with n sample units, where Y_{obs} and Y_{mis} denote the observed values and the missing values in Y, respectively. Further, $X = (X_1, ..., X_j)$ is a set of j fully observed covariates, where X_{obs} and X_{mis} correspond to the observed an missing parts in Y. We use notation n_{obs} for the number of sample units with observed values of Y and n_{mis} for the number of sample units with missing values. Finally, let R be a response indicator that is 1 if Y is observed and 0 if Y is missing.

To impute missing values in Y and to asses variances and confidence intervals for estimators based on the imputed data we use multiple imputation methods. These

methods can be described by a Bayesian approach. In case of a parametric model for the variable to be imputed, the parameters of the model are viewed as random variables to which a prior distribution is assigned. Most commonly, in this context, an uninformative prior is used. Then, taking the observed data into account, the information on the parameters is updated, leading to the posterior distribution for the parameter vector. For the monotone missing data considered here, multiple imputations for the missing values can be obtained by first drawing a value from the posterior distribution of the parameter vector and then drawing a value for each missing data point from the distribution of the missing data given the drawn value of the parameter vector and the observed data. When this procedure is repeated, say m times, m multiple imputations are obtained for each missing value that are draws from the posterior predictive distribution of the missing data.

The imputation methods discussed in the remainder of this section make use of two parametric models, the linear regression model and the logistic regression model. The linear regression model for a target variable Y can be written as

$$Y_i = X_i^T \beta + \epsilon_i,$$

with X_i the vector of values from the j covariates for unit i, β the corresponding regression coefficient vector and ϵ_i a normally distributed random error with expectation zero and variance σ^2. Parameter estimates $\hat{\beta}$, $\hat{\epsilon}_i$ and $\hat{\sigma}^2$ of this model can be obtained by ordinary least squares using the units for which both Y and X are observed. Using uninformative priors for β and σ^2 the posterior distribution for β is $N(\hat{\beta}, V(\hat{\beta}))$, i.e. normal with mean $\hat{\beta}$ and covariance matrix $V(\hat{\beta}) = \sigma^2 (X_{obs}^T X_{obs})^{-1}$ and the posterior distribution for σ^2 is given by $\hat{\epsilon}^T \hat{\epsilon}/A$, with A a chisquare variate with $n_{obs} - r$ degrees of freedom. A draw from the posterior predictive distribution for a missing value for unit i can be obtained by drawing values σ^{2*} and β^* from their posterior distributions and then drawing a value for $Y_{mis,i}$ from $N(X_i^T \beta^*, \sigma^{2*})$.

The logistic regression model for a binary (0,1) target variable W, can be expressed as

$$log \frac{\pi_i}{1 - \pi_i} = X_i^T \gamma,$$

with γ the corresponding regression coefficient vector and π_i the probability of observing $W_i = 1$ or, equivalently, $\pi_i = \mathbb{E}[W_i]$. An expression for π_i in terms of the linear predictor $X_i^T \gamma$ is obtained from the inverse logit transformation: $\pi_i = expit(X_i^T \gamma) = exp(X_i^T \gamma)/[1 + exp(X_i^T \gamma)]$. Using an uninformative prior for γ, the corresponding posterior distribution is approximately $N(\hat{\gamma}, \hat{V}(\hat{\gamma}))$ with $\hat{\gamma}$ the maximum likelihood estimator for γ and $\hat{V}(\hat{\gamma})$ the associated covariance matrix. A draw from the posterior predictive distribution of a missing value $W_{mis,i}$ can be obtained by first drawing a value γ^* from the posterior distribution for γ and then drawing a value W_i^* from a Bernoulli distribution with parameter $\pi^* = expit(X_i^T \gamma^*)$

2.2.2 Predictive mean matching

Multiply imputing Y_{mis} by means of PMM is performed by the following algorithm:

1. Use linear regression of Y_{obs} given X_{obs} to estimate $\hat{\beta}$, $\hat{\sigma}$ and $\hat{\varepsilon}$ by means of ordinary least squares.
2. Draw σ^{2*} as $\sigma^{2*} = \hat{\varepsilon}^T\hat{\varepsilon}/A$, where A is a χ^2 variate with $n_{obs}-r$ degrees of freedom.
3. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}(X_{obs}^T X_{obs})^{-1}$.
4. Calculate $\hat{Y}_{obs} = X_{obs}\ \hat{\beta}$ and $\hat{Y}_{mis} = X_{mis}\ \beta^*$.
5. For each $\hat{Y}_{mis,i}$, find $\Delta = |\hat{Y}_{obs} - \hat{Y}_{mis,i}|$.
6. Randomly sample one value from $(\Delta^{(1)}, \Delta^{(2)}, \Delta^{(3)})$, where $\Delta^{(1)}$, $\Delta^{(2)}$ and $\Delta^{(3)}$ are the three smallest elements in Δ, respectively, and take the corresponding $Y_{obs,i}$ as the imputation.
7. Repeat steps 1 through 6 m times, each time saving the completed data set.

The default of the function `mice` in the `R`-package `mice` performs multiple imputation ($m = 5$) according to the description of this algorithm. The regression function `mi.pmm` in `mi` also performs PMM imputation, but calculates $\Delta = min|\hat{Y}_{obs}-\hat{Y}_{mis,i}|$ and selects the corresponding $Y_{obs,i}$ as the imputation.

2.2.3 Two-part imputation

Let Y be decomposed into two variables $(W_i,\ Z_i)$, where Y_i denotes the ith value in Y, giving

$$W_i = \begin{cases} 1 & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0 \end{cases}. \tag{2.1}$$

$$Z_i = \begin{cases} g(Y_i) & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0 \end{cases}. \tag{2.2}$$

where g is a monotonically increasing function, chosen such that the non-zero values in Y_i are approximately normally distributed (Manning et al., 1981; Duan et al., 1983; Schafer and Olsen, 1999). Multiply imputing Y_{mis} by means of two-part multiple imputation can be done by the following algorithm as described by Schafer and Olsen (1999):

1. Use logistic regression on W_{obs} given X_{obs} to estimate $\hat{\gamma}$, $\hat{V}(\hat{\gamma})$.
2. Draw γ^* from a multivariate normal distribution centered at $\hat{\gamma}$ with covariance matrix $\hat{V}(\hat{\gamma})$.
3. Draw W_i from a Bernoulli distribution with probability $\pi_i^* = expit(X_i^T\gamma^*)$ independently for W_{mis} .
4. For all $W_i \neq 0$, use linear regression of Z_{obs} given X_{obs} to estimate the least squares estimates $\hat{\beta}$ and residuals $\hat{\varepsilon}_i = Z_i - X_i^T\hat{\beta}$ where $i \in obs$.
5. Draw a random value of σ^{2*} as $\sigma^{2*} = \hat{\varepsilon}^T\hat{\varepsilon}/A$, where A is a χ^2 variate with $n_{obs.1}-r$ degrees of freedom, with $n_{obs.1}$ the number of observed elements given $W_i = 1$.
6. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}(X_{obs}^T X_{obs})^{-1}$.

7. Draw Z_i from a normal distribution with mean $\mu_i^* = X_i^T \beta^*$ and variance σ^{2*} independently for all Z_{mis}.
8. Set $Y_i = 0$ if $W_i = 0$ and $Y_i = g^{-1}(Z_i)$ if $W_i = 1$ for all Y_{mis}.
9. Repeat the steps m times, each time saving the completed data set. Note that steps 1 and 4 do not change, and need to be done only once. Further, steps 4 through 7 are performed on the subset $W_i = 1$.

A list of software that incorporates a two-part model includes (but is not limited to) `IVEware`, `mi` and `VIM`. Note that these software packages may use different approaches to the two-part model as well as different algorithms, but all use a two-part approach. For example, `mi` log-transforms the continuous part of the data and the `VIM` routine `irmi` uses robust estimation methods.

2.2.4 Imputing through the BGLoM

The BGLoM by Javaras and Van Dyk (2003) extends the general location model (Olkin and Tate, 1961) by incorporating a two-level model. The precise model is too intricately detailed to be summarized here. Instead, well-documented EM and data augmentation algorithms can be found in Javaras and Van Dyk (2003). We use software and script, kindly provided by the authors, in our simulations.

2.3 Univariate simulation

In order to compare the performance of the imputation methods at hand, we use a design-based approach wherein we create a finite population from which we repeatedly sample. We make use of a design-based simulation because there are no statistical models that would help us generate multivariate semi-continuous data with given dependencies among the variables and fixed underlying univariate and multivariate properties. Consequently, we have chosen to generate data with known properties, and subsample from these. This procedure is popular in official statistics (see e.g., Chambers and Clark (2012); Alfons et al. (2010a,b)) and is often used in the case of performance assessment of imputation procedures in this field.

2.3.1 Generating populations

We separate the simulations on the level of the point mass and generate two populations. Both populations have size $N = 50.000$, but the populations differ in the size of the point mass: 30% and 50% point masses at zero, respectively. Note that when the size of the point mass changes, estimates such as the mean, median and variances change as well.

Step 1: Generating semicontinuous data

For each population, we start by creating a normally distributed variable $Q \sim N(5, 1)$ to which we assign a point mass at zero by drawing from a binomial distribution with a 30% (population 1) or 50% (population 2) chance for any value in Q to take on the point mass. Please note that Q is now a semicontinuous variable wherein the continuous part is normally distributed. The zeros in Q are initially completely at random, but a dependent relation with the covariate will be induced by transformation.

Step 2: Generating covariates

In order to measure the influence of the relation with the covariate, we want to create covariates with varying correlations with the simulation population Q. To do so, we defined the correlation matrix for four covariates and the semicontinuous variable Q as

$$R_{QX} = \begin{bmatrix} Q & X_1 & X_2 & X_3 & X_4 \\ 1 & & & & \\ .80 & 1 & & & \\ .50 & .4 & 1 & & \\ .30 & .24 & .15 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Using these correlations, we constructed standard deviation scores ($SDS_{X_{ij}}$), with mean zero, for the covariates according to

$$SDS_{X_{ij}} = SDS_{Q_i} * \rho_{QX_j} + \epsilon_i$$

where ρ_{QX_j} is the correlation between Q and X_j obtained from R_{QX}, SDS_{Q_i} is the standardized score of Q (with mean zero and standard deviation 1) and ϵ_i is a random draw from the normal distribution $N(0, \sqrt{1 - \rho^2_{YX_j}})$.

Step 3: Generating target variables

To create semicontinuous target variables, we used the following transformations of Q:

$$\begin{aligned} Y_1 &= Q \\ Y_2 &= Q^2/\max\{Q\} \\ Y_3 &= Q^4/\max\{Q^3\} \\ Y_4 &= Q^8/\max\{Q^7\} \\ Y_5 &= Q^{12}/\max\{Q^{11}\}, \end{aligned}$$

thereby varying the degree of skewness while keeping the variables in the same scale. For example, the continuous parts in Y_1 and Y_5 are normally distributed and extremely

skewed, respectively. Creating transformed skewed variables also introduces extreme values, which in turn may severely impair a methods imputation performance. Figure 2.1 displays histograms for Y_1 through Y_5 with a 50% point mass at zero.

Combining the set of transformed variables $Y = (Y_1, ..., Y_5)$ with the variables in $X = (X_1, ..., X_4)$ provides us with a dataset with different bivariate relations between any of the variables in Y and the covariates X. Moreover, because of the different degrees of skewness between the variables in Y, the bivariate relations between any of the variables in X and the target variables Y also differ. For example, the bivariate relations between X_1 and Y_1 are stronger than the relations between X_2 and Y_1, and the relations between X_1 and Y_2 are stronger than the relations between X_1 and Y_3.

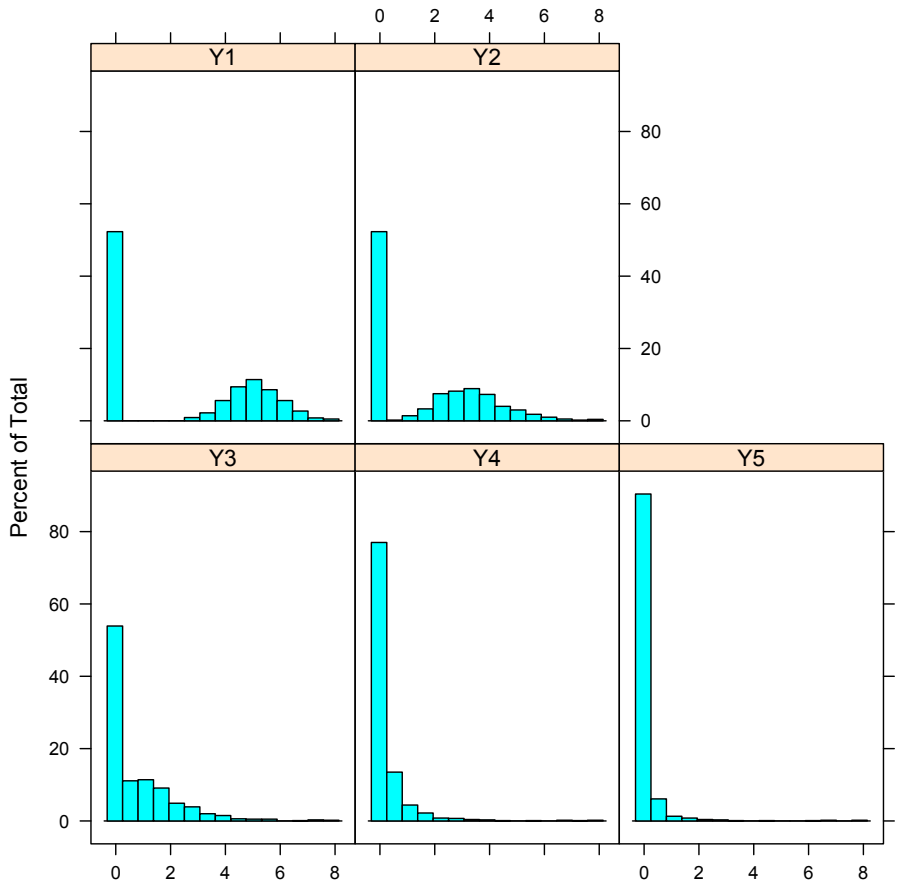

Fig. 2.1. Generated semicontinuous variables ($Y_1 - Y_5$) with a point mass at 50%

Please note that we investigate the univariate problem, meaning that we impute each of the semicontinuous variables (e.g., Y_1) based on one of the covariates (e.g., X_1).

2.3.2 Sampling from the population

To investigate the performance of the methods under different sample sizes, we randomly sample from the combined set of Y and X for each population. We used samples of size 100, 500 and 1000, respectively. Other sampling schemes are beyond the scope of this research, because we are mainly interested in the missing data process and not in the sampling process.

2.3.3 Generating missingness

Because we investigate the univariate case, we may impose the missingness for each sample in all Y simultaneously. We created missingness in our samples according to the following missing at random (MAR) mechanism:

$$P(R = 0|Y_{obs}, Y_{mis}, X_j) = P(R = 0|Y_{obs}, X_j)$$

by using a random draw from a binomial distribution of the same length as Y and of size 1 with missingness probability equal to the inverse logit

$$P(R = 0) = \frac{e^a}{(1 + e^a)}.$$

In the case of left-tailed MAR missingness, $a = (-\bar{X}_j + X_{ij})/\sigma_{X_j}$ gives 50% missingness, where σ_{X_j} indicates the standard deviation of variable X_j. For right-tailed MAR missingness, this can be achieved by choosing $a = (\bar{X}_j - X_{ij})/\sigma_{X_j}$. Choosing $a = .75 - [(\bar{X}_j - X_{ij})/\sigma_{X_j}]$, or $a = -.75 + [(\bar{X}_j - X_{ij})/\sigma_{X_j}]$, gives 50% centered MAR missingness or 50% tailed MAR missingness, respectively. Adding or subtracting a constant moves the sigmoid curve, which results in different missingness proportions.

The samples in which missingness was imposed, were imputed and evaluated. Separate simulations were done for 25% and 50% missingness per variable. All simulations have been carried out in R 2.13 and are repeated 100 times. The function `mice(data, method = "pmm")` from the R-package `mice` (version 2.13) (Van Buuren and Groothuis-Oudshoorn, 2011) was used for PMM.

A custom adaptation of `mice` was developed for two-part imputation which uses `mice(data)` with method specification `method = "logreg"` for the binary indicator and `method = "norm"` for the continuous part. After the final iteration of the algorithm a post processing command is parsed which sets all zeros from the imputed binary indicator to zeros in the continuous data. The function `mi()` from the *R*-package `mi` (version 0.09-18) was used to impute the object which has been pre-processed by the function `mi.preprocess(data)`. Finally, the function `irmi(data)` with semicontinuous columns indicated as `mixed` from the *R*-package `VIM` (version 3.0.1) was used for imputations based on the `irmi` algorithm.

2.3.4 Evaluation of imputations

In the case of a simulated dataset, evaluations can be done because 'truth' is known. In case of a real-life dataset, containing observed missingness, this cannot be done, because the actual values are unknown. It is therefore necessary to check the imputations in real-life datasets by means of a standard of reasonability: differences between observed and imputed values and distributional shapes can be checked to see whether they make sense given the particular dataset (see Abayomi et al. (2008) for more information on this subject).

We evaluate the quality of imputations by assessing the following criteria: bias of the mean, median and correlation, coverage of the 95% confidence interval of the mean, the size of the point mass, preservation of distributional shapes, and the plausibility of the imputed data. We asses plausibility by looking whether the imputed values are realistic given the observed data, For example, could they have been observed if the data were not missing.

2.4 Univariate Results

2.4.1 Bias of the mean

Tables 2.1 and 2.2 display biases in the mean for Y_1 through Y_2 after imputation given the covariates X_1 and X_4, respectively. Bias of the mean is defined as the difference between the recovered mean and the population mean. From these tables, it can be seen that PMM and the two-part model estimate the mean very accurately. The bias from the population mean for these methods is very low, regardless of the varying simulation conditions. However, the BGLoM, `mi` and `irmi` seem somewhat biased in certain cases. The bias of the BGLoM depends on the missingness mechanism and is especially visible in the case of left-tailed MAR-missingness. Also, observe that the bias depends on the size of the point mass. It seems that the BGLoM overestimates the smaller point masses, thereby making the data more semicontinuous than it should be. Especially when combined with a 'weaker' covariate, mean biases for the BGLoM become much larger when the size of the point mass decreases.

The bias of `mi` is larger for right-tailed and left-tailed MAR-missingness, although this difference disappears when the variable becomes more skewed. For the non-correlating covariate (X_4), all biases for `mi` are very small.

In contrast, the bias of the mean for `irmi` is very small for a high-correlating covariate but very large for the non-correlating covariate.

For all methods, the absolute bias decreases when the variable with missingness become skewed, that is, for Y_2 through Y_5. This, however, is to be expected, because with more-skewed variables, means and variances are closer to zero than in the case of less-skewed variables (see Figure 2.1). For all three methods, bias increases with the percentage of missingness, but this effect is much more pronounced for the BGLoM and for `mi`. The bias of the mean of `mi` for simulations with less (25%) missingness is comparable to the bias of the mean of PMM (not shown).

2.4.2 Bias of the correlation with the covariate

Figure 2.2 displays the difference between the true correlation and the recovered correlation (correlation bias). Correlation bias is smaller for PMM, `irmi` and the two-part model, than for the BGLoM, even for skewed semicontinuous variables. However, when variables become more skewed (e.g., in the case of Y_4 and Y_5), correlations for PMM and the two-part model tend to be overestimated. `irmi` correlations are always overestimated. The amount of overestimation increases for variables that are more skewed. PMM, `irmi` and the two-part model are clearly sensitive to extreme skewness, for example, in Y_4 and Y_5.

`mi` produces large correlation bias even in the case of Y_1 and there does not seem to be any relation to the missingness mechanisms. For `mi` it shows that the combination between skewed data and tailed MAR missingness systematically results in large correlation bias. For Y_5 we note that besides the much larger bandwidth, the maximum bias of the correlation for `mi` is smaller than the maximum bias of any other method.

The results and findings are similar for the uncorrelated covariate X_4 (not shown). For all three methods, it holds that correlation biases become smaller when sample

Table 2.1. Univariate simulation results for X_1 over 100 simulations. The table depicts bias of the mean, coverage rate for the mean, CI width and the estimated percentage of zeros obtained using different imputation methods and different missingness mechanisms for semicontinuous variables Y_1 through Y_3. All cases represent a sample size of n=500 and 50% MAR missingness.

			2-Part				BGLoM				PMM				MI				IRMI			
	pm	mar	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero
Y_1	0.3	Left	0.00	0.96	0.55	0.30	-0.38	0.20	0.56	0.38	-0.01	0.97	0.55	0.30	0.19	0.88	0.81	0.26	-0.04	0.79	0.44	0.31
	0.3	Mid	-0.04	0.95	0.55	0.31	-0.08	0.97	0.58	0.32	-0.01	0.94	0.51	0.30	-0.05	0.91	0.62	0.32	0.07	0.86	0.43	0.29
	0.3	Right	-0.02	0.96	0.53	0.31	0.05	0.99	0.80	0.29	-0.01	0.93	0.49	0.30	-0.10	0.84	0.60	0.32	0.03	0.91	0.43	0.29
	0.3	Tail	0.02	0.95	0.50	0.30	-0.13	0.91	0.58	0.33	-0.02	0.94	0.49	0.30	0.03	0.95	0.60	0.29	-0.02	0.90	0.44	0.31
	0.5	Left	0.03	0.99	0.55	0.49	-0.19	0.79	0.53	0.54	0.00	0.95	0.53	0.50	0.13	0.93	0.67	0.47	-0.09	0.79	0.46	0.52
	0.5	Mid	0.03	0.93	0.58	0.49	0.01	0.94	0.61	0.50	0.00	0.89	0.54	0.50	0.01	0.95	0.65	0.50	0.02	0.82	0.46	0.50
	0.5	Right	-0.02	0.95	0.59	0.50	0.13	0.96	0.95	0.46	0.00	0.95	0.55	0.50	-0.19	0.85	0.79	0.53	0.04	0.89	0.45	0.49
	0.5	Tail	-0.01	0.95	0.52	0.50	0.01	0.90	0.58	0.50	0.00	0.97	0.52	0.50	-0.03	0.94	0.58	0.50	-0.01	0.94	0.46	0.50
Y_2	0.3	Left	0.00	0.95	0.35	0.30	-0.20	0.34	0.34	0.38	-0.01	0.91	0.34	0.30	0.10	0.90	0.44	0.26	-0.02	0.90	0.29	0.31
	0.3	Mid	-0.03	0.96	0.36	0.31	-0.04	0.97	0.40	0.32	0.00	0.97	0.34	0.30	-0.01	0.95	0.43	0.32	0.04	0.86	0.28	0.29
	0.3	Right	-0.02	0.94	0.37	0.31	-0.01	0.96	0.68	0.29	0.00	0.91	0.36	0.30	-0.06	0.91	0.51	0.32	0.01	0.84	0.27	0.29
	0.3	Tail	0.00	0.96	0.33	0.30	-0.07	0.91	0.39	0.33	-0.02	0.92	0.33	0.30	0.02	0.95	0.44	0.29	-0.01	0.90	0.28	0.31
	0.5	Left	0.02	0.98	0.33	0.49	-0.10	0.83	0.32	0.54	0.00	0.98	0.33	0.50	0.06	0.93	0.42	0.47	-0.04	0.80	0.29	0.52
	0.5	Mid	0.01	0.97	0.36	0.49	0.00	0.91	0.38	0.50	0.00	0.90	0.34	0.50	0.02	0.92	0.41	0.50	0.02	0.88	0.29	0.50
	0.5	Right	-0.01	0.98	0.40	0.50	0.05	1.00	0.84	0.46	0.00	0.96	0.38	0.50	-0.13	0.76	0.53	0.53	0.01	0.87	0.27	0.49
	0.5	Tail	-0.01	0.94	0.35	0.50	0.00	0.99	0.41	0.50	0.00	0.94	0.34	0.50	-0.02	0.95	0.41	0.50	0.00	0.94	0.28	0.50
Y_3	0.3	Left	0.00	0.94	0.18	0.30	-0.06	0.69	0.17	0.38	0.00	0.94	0.17	0.30	0.02	1.00	0.11	0.26	-0.01	0.96	0.15	0.31
	0.3	Mid	-0.01	0.96	0.20	0.31	-0.01	0.99	0.25	0.32	0.00	0.97	0.18	0.30	0.02	0.99	0.16	0.32	0.01	0.91	0.15	0.29
	0.3	Right	-0.02	0.85	0.20	0.31	-0.01	1.00	0.55	0.29	0.00	0.96	0.22	0.30	0.03	0.99	0.31	0.32	-0.02	0.76	0.13	0.29
	0.3	Tail	-0.01	0.94	0.18	0.30	-0.02	0.99	0.22	0.33	-0.01	0.86	0.18	0.30	0.03	0.99	0.20	0.29	-0.01	0.85	0.14	0.31
	0.5	Left	0.00	0.98	0.16	0.49	-0.03	0.82	0.14	0.54	0.00	0.96	0.15	0.50	0.01	0.96	0.07	0.47	-0.01	0.88	0.13	0.52
	0.5	Mid	0.01	0.93	0.18	0.49	0.00	0.97	0.21	0.50	0.00	0.94	0.16	0.50	0.02	0.97	0.13	0.50	0.01	0.87	0.14	0.50
	0.5	Right	-0.01	0.88	0.20	0.50	0.00	1.00	0.78	0.46	0.00	0.86	0.19	0.50	0.02	0.97	0.25	0.53	-0.02	0.74	0.11	0.49
	0.5	Tail	-0.01	0.91	0.17	0.50	0.01	1.00	0.27	0.50	0.00	0.92	0.17	0.50	0.02	0.99	0.21	0.50	-0.01	0.88	0.13	0.50

size increases, but there is no clear relation with the size of the point mass and the amount of missingness.

2.4.3 Bias of the median

Estimating the median for Y_1 through Y_5 from imputed data can lead to large biases, especially when the population has been randomly assigned 49% of zeros and the imputed data returns 51% of zeros. Biases of the median are therefore mostly influenced by the size of the point mass, with biases being much lower for data with 30% zeros. Besides, when skewness increases in the simulation data, point estimates move closer to zero, resulting in biases being very near to zero for Y_4 and Y_5 for all methods (see Figure 2.3).

In all other cases, PMM and the two-part model are less biased than `mi`, `irmi` and the BGLoM. Further, the spread in the biases is much lower for PMM than for `irmi`, `mi` and the BGLoM, but is similar between PMM and the two-part model. The amount of missingness does not influence the extent of the bias, neither does the missingness mechanism, nor does the sample size. The non-correlating covariate results in slightly smaller median biases for all methods.

Table 2.2. Univariate simulation results for X_4 over 100 simulations. The table depicts bias of the mean, coverage rate for the mean, CI width and the estimated percentage of zeros obtained using different imputation methods and different missingness mechanisms for semicontinuous variables Y_1 through Y_3. All cases represent a sample size of n=500 and 50% MAR missingness.

			2-Part				BGLoM				PMM				MI				IRMI			
	pm	mar	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero	bias	cov	ciw	zero
Y_1	0.3	Left	0.02	0.98	0.81	0.30	-0.51	0.59	1.15	0.40	0.00	0.97	1.38	0.30	-0.01	0.94	0.88	0.31	0.69	0.00	0.36	0.15
	0.3	Mid	0.00	0.96	0.69	0.30	-0.48	0.63	1.07	0.40	0.01	0.91	1.21	0.30	0.00	0.97	0.81	0.30	0.70	0.00	0.36	0.16
	0.3	Right	-0.01	0.92	0.76	0.31	-0.60	0.39	1.11	0.42	0.03	0.88	1.47	0.30	0.04	0.98	0.93	0.30	0.68	0.00	0.36	0.15
	0.3	Tail	0.01	0.94	0.70	0.30	-0.50	0.55	1.03	0.40	0.05	0.93	1.08	0.29	0.01	0.94	0.85	0.30	0.73	0.00	0.36	0.15
	0.5	Left	0.01	0.96	0.83	0.50	0.05	1.00	1.08	0.49	-0.05	0.91	1.53	0.51	0.00	0.96	0.99	0.50	-0.20	0.00	0.40	0.53
	0.5	Mid	-0.01	0.93	0.73	0.50	-0.02	0.96	1.00	0.50	-0.01	0.94	1.54	0.50	0.00	0.90	0.76	0.50	0.15	0.03	0.40	0.47
	0.5	Right	0.01	0.91	0.82	0.50	0.01	1.00	1.32	0.49	-0.03	0.89	1.73	0.50	-0.01	0.94	0.97	0.50	-0.03	0.00	0.40	0.50
	0.5	Tail	0.01	0.96	0.72	0.50	-0.01	1.00	1.09	0.50	0.03	0.88	1.22	0.49	0.00	0.96	0.87	0.50	-0.02	0.49	0.43	0.50
Y_2	0.3	Left	0.02	0.96	0.52	0.30	-0.30	0.59	0.66	0.40	-0.02	0.89	0.83	0.31	-0.01	0.94	0.60	0.31	0.34	0.00	0.26	0.15
	0.3	Mid	0.00	0.94	0.43	0.30	-0.27	0.52	0.63	0.40	-0.02	0.89	0.84	0.30	0.00	0.94	0.54	0.30	0.38	0.00	0.26	0.16
	0.3	Right	0.00	0.95	0.50	0.31	-0.34	0.39	0.65	0.42	0.01	0.91	0.99	0.30	0.03	0.95	0.67	0.30	0.34	0.01	0.26	0.15
	0.3	Tail	0.00	0.90	0.44	0.30	-0.30	0.57	0.63	0.40	0.03	0.93	0.80	0.29	0.01	0.98	0.55	0.30	0.39	0.00	0.27	0.15
	0.5	Left	0.00	0.98	0.52	0.50	0.03	1.00	0.71	0.49	-0.01	0.89	0.90	0.50	0.00	0.96	0.59	0.50	-0.13	0.00	0.25	0.53
	0.5	Mid	-0.01	0.92	0.45	0.50	-0.02	0.92	0.59	0.50	0.01	0.91	1.04	0.50	0.00	0.94	0.56	0.50	0.07	0.03	0.26	0.47
	0.5	Right	0.01	0.93	0.52	0.50	-0.01	0.98	0.80	0.49	-0.02	0.92	0.97	0.50	0.01	0.97	0.62	0.50	-0.03	0.00	0.26	0.50
	0.5	Tail	0.00	0.95	0.44	0.50	-0.03	1.00	0.61	0.50	0.02	0.91	0.72	0.49	0.02	0.97	0.57	0.50	-0.01	0.50	0.27	0.50
Y_3	0.3	Left	0.01	0.95	0.25	0.30	-0.10	0.60	0.29	0.40	0.00	0.91	0.42	0.30	0.02	0.94	0.34	0.31	0.09	0.38	0.15	0.15
	0.3	Mid	0.00	0.93	0.21	0.30	-0.09	0.66	0.26	0.40	-0.02	0.92	0.37	0.31	0.01	0.95	0.31	0.30	0.11	0.24	0.15	0.16
	0.3	Right	0.00	0.95	0.25	0.31	-0.12	0.55	0.27	0.42	0.01	0.93	0.39	0.29	0.03	0.93	0.36	0.30	0.09	0.36	0.15	0.15
	0.3	Tail	0.00	0.97	0.22	0.30	-0.11	0.56	0.27	0.40	0.01	0.92	0.40	0.30	0.02	0.91	0.28	0.30	0.12	0.20	0.15	0.15
	0.5	Left	0.00	0.94	0.23	0.50	0.02	1.00	0.25	0.49	-0.02	0.90	0.37	0.50	0.01	0.95	0.31	0.50	-0.06	0.02	0.12	0.53
	0.5	Mid	0.00	0.90	0.19	0.50	-0.01	0.92	0.29	0.50	-0.01	0.93	0.39	0.50	0.01	0.93	0.28	0.50	0.01	0.06	0.12	0.47
	0.5	Right	0.01	0.93	0.23	0.50	0.00	0.98	0.45	0.49	-0.01	0.97	0.46	0.50	0.02	0.96	0.38	0.50	-0.02	0.00	0.12	0.50
	0.5	Tail	0.00	0.95	0.21	0.50	-0.02	0.99	0.27	0.50	0.00	0.88	0.29	0.49	0.03	0.96	0.30	0.50	-0.01	0.46	0.13	0.50

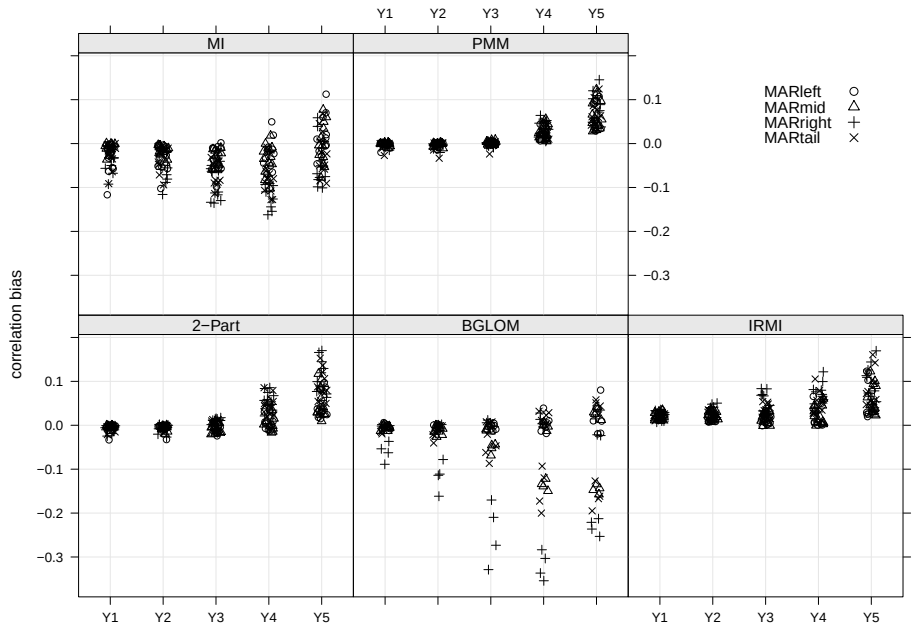

Fig. 2.2. Bias of the correlation with the covariate X_1 for different imputation methods over 100 simulations.

2.4.4 Coverage rates and confidence interval widths

PMM and the two-part model have very consistent coverages, whereas `irmi` and BGLoM coverages tend to vary to a great extent. `mi` shows a pattern opposite to that of PMM and the two-part model. With MI, increasingly skewed variables show increasingly higher coverages. The same holds for the BGLoM, but to a much lesser extent. The BGLoM and `mi`, occasionally, even display a 100% coverage over 100 simulations (see Figure 2.4). However, we can see in Figure 2.5 that `mi` and the BGLoM also have much wider confidence intervals. This only holds for covariates that have predictive power.

When there is no relation with the covariate (e.g., as in X_4), the two-part model shows the smallest confidence interval widths with consistent coverages. BGLoM and `mi` confidence interval widths are also smaller than the confidence interval widths for PMM, although this difference disappears as variables become more skewed. More, PMM coverages for data with a 30% point mass are much higher than BGLoM coverages in the case of low-correlating predictors.

The `irmi` algorithm shows a severe problem: confidence interval widths are small, but coverage rates are either 0 or very small. This only happens in the case of a single

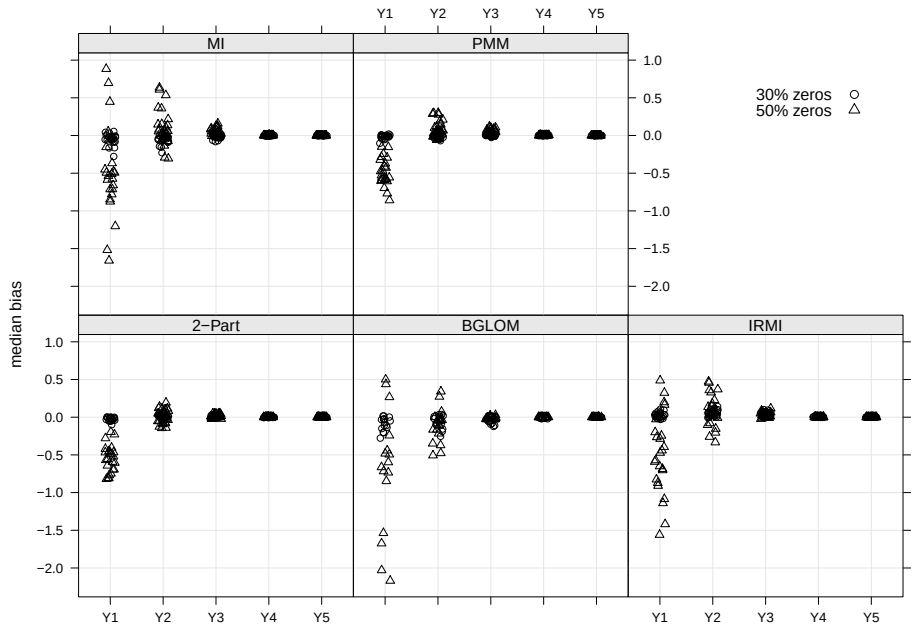

Fig. 2.3. Bias of the median for different sizes of the point mass over 100 simulations given covariate X_1.

non-correlating covariate. As soon as there is some predictive power, results improve, although the coverage rates are never on par with PMM or `mi`. The reason for this phenomenon is the logistic step in the algorithm either appoints the missing data as continuous or as part of the point mass, resulting in 75% or 25% zeros (in the case of a 50% point mass at zero with 50% missingness. The average of all imputed means over 100 simulations may be close to the population mean, but the confidence intervals of those respective means do not contain the population mean. PMM and the two-part model show lower coverages for missingness mechanisms that involve the right tail of the data, but only for Y_4 and Y_5, where skewness moves to the extreme. `irmi` also displays this trend, but to a much greater extent. The BGLoM, on the other hand, shows unacceptable coverages for left-tailed missingness, but this trend weakens when skewness moves to the extreme. For right-tailed missingness, the BGLoM and `mi` show larger confidence interval widths, whereas the confidence interval widths for PMM and for the two-part model are not clearly influenced by the location of the missingness. Please note that for PMM, `irmi` and the two-part model it can be clearly seen that for each variable there are three clusters of points. These clusters correspond to the three sample sizes, where the smaller sample sizes result in larger confidence interval widths.

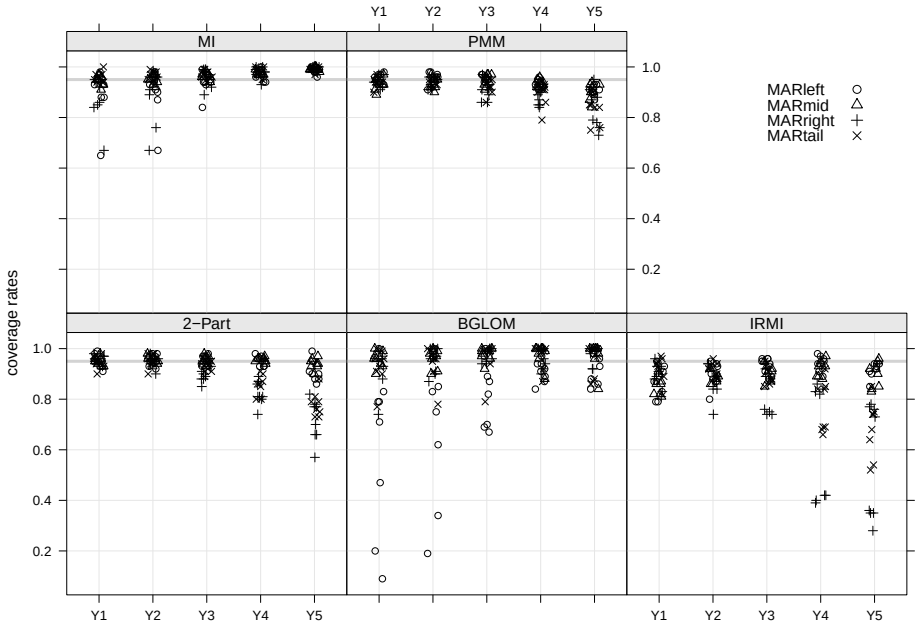

Fig. 2.4. Coverage rates for different imputation methods over 100 simulations using covariate X_1.

In general, when there is at least some predictive power, PMM coverage rates and confidence interval widths outperform those of `irmi`, `mi` and the BGLoM. Further, two-part and PMM coverage rates and confidence intervals are very similar, with PMM having less variation between the different MAR mechanisms.

Lower percentages of missingness result in (slightly) higher coverage rates, as do larger sample sizes.

2.4.5 Point mass

Table 2.1 and Table 2.2 also show the percentage of estimated amount of zeros (point mass), for the simulated conditions. The performance of PMM and the two-part model does not rely on the size of the point mass. Both algorithms estimate the size of the point mass correctly, with very small deviations, regardless what the simulation conditions are. See Figure 2.6 for a graphical representation of the biases of the estimated point mass. The BGLoM, on the other hand, is not that insensitive against the size of the point mass, as we have already seen in previous paragraphs. For the BGLoM, the estimation of the amount of zeros heavily depends on the size of the point mass in the original data and the missingness mechanism.

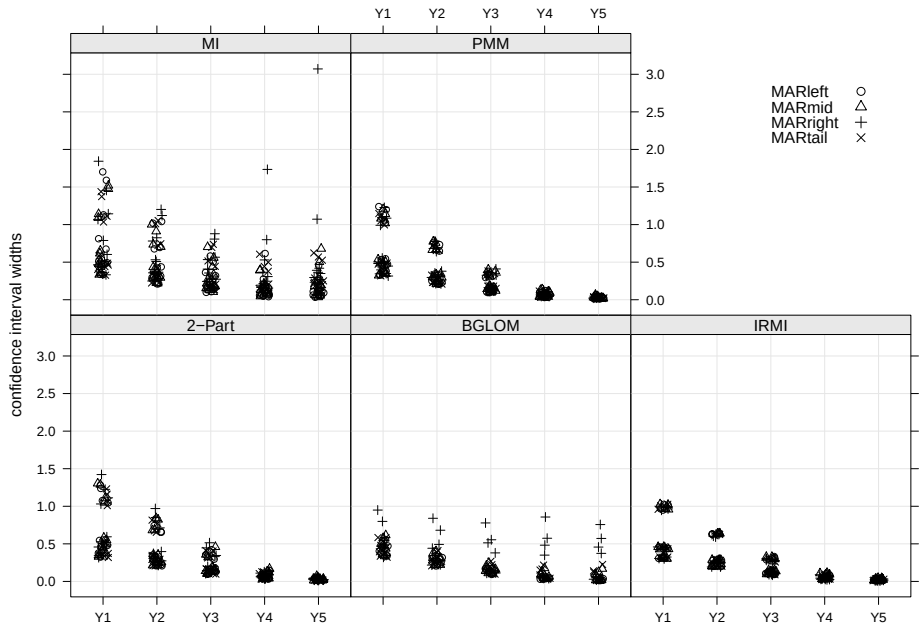

Fig. 2.5. Confidence interval widths for different imputation methods over 100 simulations using covariate X_1.

The point mass estimated by `mi` is acceptable in the case of a high-correlating covariate, although PMM, `irmi` and the two-part model are more accurate. In the case of a non-correlating covariate, the amount of zeros estimated by `mi` is comparable to PMM and the two-part model.

As we have mentioned in Section 2.4.4, in the case of a single non-correlating covariate, `irmi` performance could be improved. For the 50% point mass, the average amount of zeros is very close to the population point mass; however, the individual point masses are either 25% or 75%. For the 30% point mass this biased estimation of the zeros becomes more apparent. Table 2.2 shows this underestimation of the 30% point mass by the `irmi` algorithm.

The estimation of the zeros by `irmi` also differs from the other methods with a two-stage approach. The amount of zeros and the location of the zeros is the same for each of the m multiply imputed datasets, meaning that there is less between imputation variance than multiple imputation theory dictates. This can be easily solved by drawing β^* for each of the m multiple imputation streams from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\hat{V}(\hat{\beta})$, conform the algorithm in Section 4.2.2.

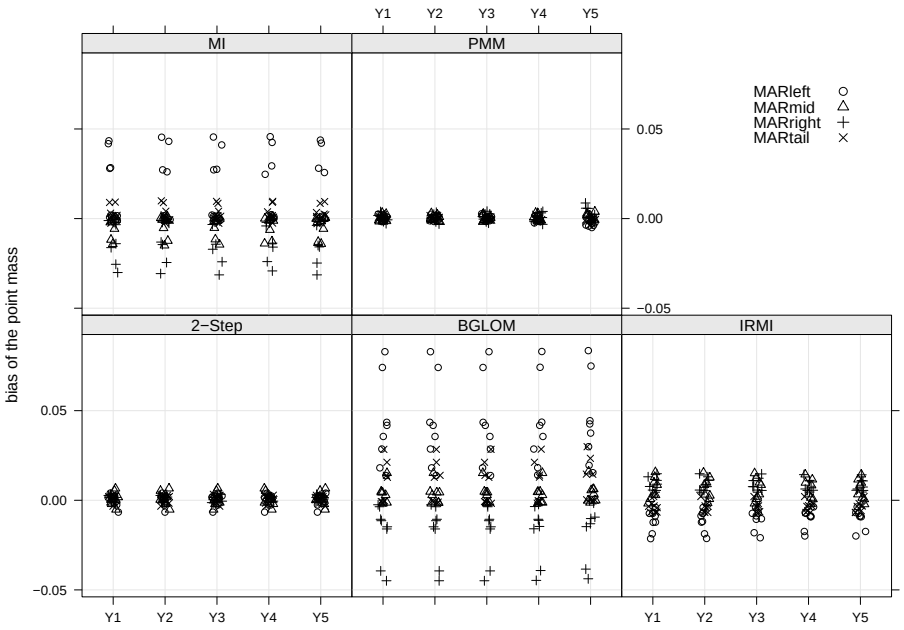

Fig. 2.6. Bias of the estimated size of the point mass for different imputation methods over 100 simulations using covariate X_1.

The amount of skewness does not influence the bias of the point mass estimate. Please note that point masses for `mi`, `irmi`, the two-part model and the BGLoM are equal for matching simulation conditions on different variables. This is due to the sequential nature of these methods, where the imputation of zeros is treated fixed.

2.4.6 Distributional shapes

PMM preserves the distributional shapes of the variables, even for the most extremely skewed semicontinuous variables, although some information is lost in the right tail of the distributions due to sampling. `mi` imputes a log-transformation of the continuous part of a semicontinuous variable, which clearly shows from the plots. Non-negative data are imputed and the larger part of the imputations follows the original data distribution. However, medians are underestimated and extreme values are imputed on the right-tail side, because the back transformation of the log-transformed data introduces extreme imputed values.

`irmi` imputations produce imputations that are similar to the original data distribution, but only for Y_1 and Y_2. As variables become more skewed, distributions of completed data become very similar to those of the two-part model. For the BGLOM,

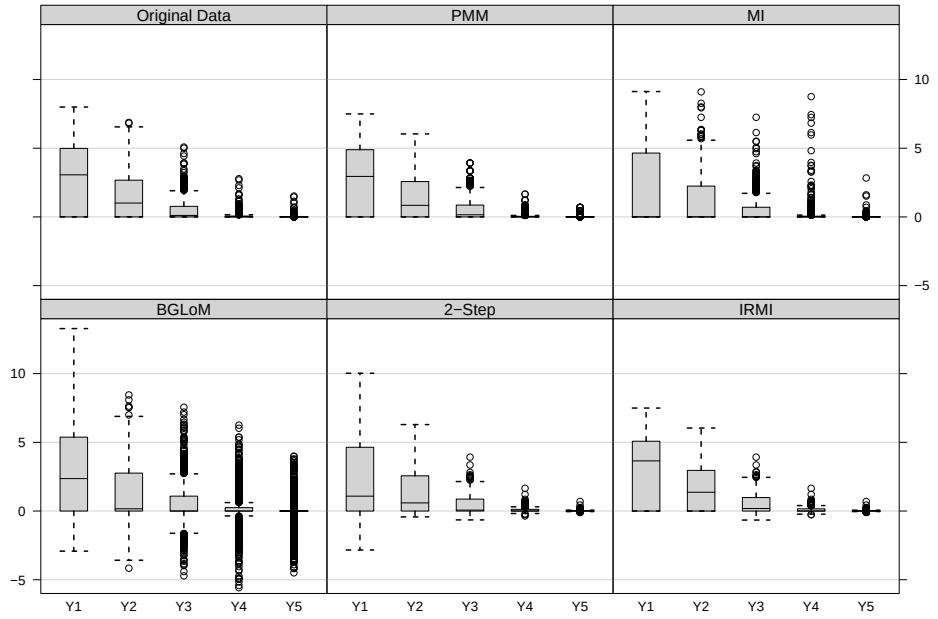

Fig. 2.7. Right-tailed MAR missingness: Boxplots of the original data and imputed data for 5 imputation methods for 50% missing data. Imputations are based on covariate X_1.

two-part imputation and `irmi`, it holds that when skewness increases, these model-based methods tend to represent a normal curve again (see Figure 2.7).

2.4.7 Plausibility of the imputations

The original data are non-negative, but the two-part model, `irmi` and the BGLoM will also impute negative values. In contrast, PMM and `mi` will impute only positive data, thus closer resembling the original distribution. However, `mi` imputes implausible values in the right tail, moving outside the range of population values. The hot-deck nature of PMM prevents imputations from moving outside the range of observed values, thus preserving the data distribution in this respect. This is a particular useful feature if the original data distributions and relations are to be preserved for further analysis.

2.5 Multivariate simulation

In order to be able to compare the performance of the imputation methods under multivariate missingness, we create a population from which we sample. Just like

the univariate situation, the population has size $N = 50.000$, but we fix the point mass to a 50% point mass at zero. We used simple random samples of size 1000. We consider multivariate simulations under a normal distribution, simulations for skewed distributions, and simulations for skewed distributions with outliers.

2.5.1 Generating semicontinuous population data

We aim to create a population with two semicontinuous variables Y_1 and Y_2 and a covariate X where all three variables are correlated. To this end, we start by creating two normally distributed variables $Q_1 \sim N(5, 1)$ and $Q_2 \sim N(5, 1)$ to which we assign a point mass at zero by drawing from a binomial distribution with a 50% chance for any value in Q_1 or Q_2 to take on the point mass. Please note that the results are again two semicontinuous variables wherein the continuous part is normally distributed. For the normal multivariate simulation, we set

$$T_1 = Q_1$$
$$T_2 = Q_2,$$

and for the multivariate simulation with skewed variables and with outliers, we use the following transformations:

$$T_1 = Q_1^4/\max\{Q_1^3\}$$
$$T_2 = Q_2^4/\max\{Q_2^3\},$$

and we create a covariate $W \sim N(5, 1)$ independent of the other variables. These three variables can be combined in a data matrix $D = [T_1 T_2 W]$. By construction, the three variables T_1, T_2 and W are uncorrelated. To introduce correlation, we specify the following target correlation matrix:

$$R_{YX} = \begin{bmatrix} Y_1 & Y_2 & X \\ 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}.$$

Now we find a matrix U such that $U^T U = R_{YX}$ and we transform T_1, T_2 and W to the final correlated variables by transforming the data matrix D to the final data matrix D_c by $D_c = [Y_1 Y_2 W] = DU$. Any 'transformed' zeros in Y_2 are set to zero. The following cross table shows the partitioning of the data in four parts. Within

	$Y_2 = 0$	$Y_2 \neq 0$
$Y_1 = 0$	A (0.250)	C (0.252)
$Y_1 \neq 0$	B (0.249)	D (0.249)

brackets are cross-tabulated proportions of the point mass and continuous parts of both variables as observed in the population.

Table 2.3. Normal simulations: Biases and coverage rates for the mean of the multivariate normal simulation. All biases depict the average simulation value subtracted by the population value. Please note that the bias in A, B, C and D are observed proportions minus true proportions.

		A	B	C	D	$\bar{Y}_1$	cov$\bar{Y}_1$	$\bar{Y}_2$	cov$\bar{Y}_2$	ρ_{Y_1,Y_2}
	mcar	0.001	0.000	-0.001	-0.001	-0.003	0.963	-0.008	0.945	-0.002
	left	-0.190	0.041	-0.082	0.231	1.486	0.000	1.312	0.000	-0.076
CCA	right	0.230	-0.082	0.037	-0.185	-1.406	0.000	-1.160	0.000	-0.182
	tail	-0.075	0.068	0.077	-0.070	-0.086	0.894	-0.178	0.789	-0.312
	mid	0.073	-0.071	-0.077	0.075	0.114	0.931	0.234	0.815	0.261
	mcar	0.023	-0.023	-0.025	0.026	0.004	0.937	0.014	0.947	-0.006
	left	0.040	-0.019	-0.041	0.021	0.003	0.960	-0.016	0.957	0.027
PMM	right	0.011	-0.028	-0.008	0.026	-0.024	0.953	-0.061	0.905	-0.062
	tail	0.019	-0.020	-0.019	0.022	-0.002	0.952	-0.055	0.922	-0.029
	mid	0.034	-0.030	-0.038	0.034	0.009	0.940	0.038	0.946	0.030
	mcar	0.030	-0.028	-0.029	0.028	-0.015	0.965	0.013	0.957	0.066
	left	0.030	-0.028	-0.029	0.028	-0.012	0.947	0.013	0.958	0.071
2-Part	right	0.028	-0.028	-0.027	0.029	-0.014	0.948	0.017	0.952	0.057
	tail	0.018	-0.016	-0.017	0.016	-0.016	0.940	-0.006	0.955	0.028
	mid	0.044	-0.039	-0.044	0.042	-0.006	0.944	0.025	0.946	0.106
	mcar	0.059	-0.055	-0.059	0.056	-0.009	0.950	-0.028	0.936	0.143
	left	0.041	-0.041	-0.067	0.069	0.128	0.818	0.055	0.937	0.131
MI	right	0.064	-0.057	-0.041	0.035	-0.173	0.725	-0.210	0.742	0.077
	tail	0.040	-0.036	-0.037	0.034	-0.049	0.920	-0.119	0.863	0.058
	mid	0.073	-0.065	-0.075	0.069	0.016	0.947	0.013	0.949	0.189
	mcar	-0.002	0.004	0.001	-0.002	0.016	0.838	0.020	0.755	-0.029
	left	0.021	0.008	-0.017	-0.011	-0.005	0.788	-0.113	0.624	-0.008
IRMI	right	-0.011	0.015	-0.006	0.004	0.084	0.704	0.002	0.489	-0.032
	tail	0.016	-0.009	-0.016	0.011	0.010	0.898	0.017	0.764	0.026
	mid	-0.017	0.018	0.013	-0.013	0.028	0.734	0.003	0.632	-0.071
	mcar	0.024	-0.009	-0.020	0.006	-0.033	1.000	-0.104	1.000	-0.023
	left	-0.017	-0.008	-0.017	0.043	0.179	1.000	0.115	1.000	-0.001
BGLoM	right*	-0.004	0.009	-0.002	-0.002	0.005	1.000	-0.017	1.000	-0.030
	tail	0.004	0.010	0.008	-0.020	-0.133	1.000	-0.152	0.930	-0.101
	mid	0.037	-0.035	-0.037	0.036	-0.006	0.941	0.014	1.000	0.056

* BGLoM right-tailed missingness was simulated with 25% missingness because of algorithmic problems with large amount of right-tailed missingness for normally distributed continuous parts

We create multivariate missingness following the procedure as described in Section 2.3.3 with difference that missingness in each Y is not imposed for all Y simultaneously but depends on the other variables in the data.

For the multivariate simulation with outliers, the preceding procedure is used to create an additional 500 values with $Q_1 \sim N(7,1)$ and $Q_2 \sim N(7,1)$, leading to an outlier percentage of approximately 1% in each drawn sample.

2.6 Multivariate results

2.6.1 Multivariate normal

The results of the multivariate normal simulations can be found in Table 2.3. All investigated methods retrieve the correct proportions for cells A, B, C and D, with the exception of complete case analysis (CCA). `mi` proportions seem somewhat more biased than proportions for other methods.

The same results can be found for the correlation between the two semicontinuous variables. All imputation approaches retrieve this correlation with low bias, but `mi` seems to struggle with missing completely at random (MCAR) already. This is due to `mi` log-transforming all incomplete semicontinuous data before imputation, even when the continuous parts follow a normal distribution.

PMM and the two-step method performed well as biases of the means of Y_1 and Y_2 are low, their coverage rates are acceptable and plausible and the correlation between Y_1 and Y_2 is accurately retrieved. The correlation bias for the two-step method is rather large for missingness mechanisms that involve the middle of the data.

`irmi` performance is good, for all estimates except the coverage of the mean. This indicates that `irmi` does not include enough between variation in the imputations when used as a multiple imputation approach.

The BGLoM performs well for all measures, except for the correlation between Y_1 and Y_2 for tailed missingness. Also, biases for Y_1 and Y_2 are quite large in situations where the missingness mechanism involves the left tail. Maybe coverage of the mean of Y_1 and Y_2 is a bit too well, as coverage rates tend to be 1. Comparing these results with those for the univariate case shows that the BGLoM clearly benefits from the multivariate nature of the data.

2.6.2 Multivariate skewed

It is known that some of the tested methods rely on symmetry. As a remedy, appropriate transformations could be used to transform skewed data accordingly. However, we find the skewed data case itself still of interest. As seen in the univariate simulations, back-transforming data may lead to imputing extreme values. Also, a log-transformation may not always be the most appropriate transformation for the whole data, making transforming the data a potentially tedious job, thereby delaying the imputation stage. Performance assessment of a method for imputing skewed semicontinuous data that does not necessarily require a transformation, such as PMM, is therefor still useful. The results of the multivariate simulation with skewed target variables can be found in Table 2.4.

All investigated methods retrieve the correct proportions for cells A, B, C and D, with the exception of `irmi`. Applying a log-transformation to the incomplete data before imputing with `irmi`, led to a minor decrease in performance. Because of this, we decided to post the results for `irmi` without using a transformation.

Table 2.4. Skewed simulations: Biases and coverage rates for the mean of the multivariate skewed simulation. All biases depict the average simulation value subtracted by the population value. Please note that the bias in A, B, C and D are observed proportions minus true proportions.

		A	B	C	D	$\bar{Y}_1$	cov$\bar{Y}_1$	$\bar{Y}_2$	cov$\bar{Y}_2$	ρ_{Y_1,Y_2}
	mcar	0.001	0.000	-0.001	-0.001	-0.001	0.957	-0.002	0.950	-0.006
	left	-0.075	0.018	-0.020	0.077	0.210	0.008	0.182	0.062	-0.004
CCA	right	0.073	-0.019	0.014	-0.068	-0.169	0.002	-0.152	0.018	-0.062
	tail	0.000	0.002	0.008	-0.010	-0.049	0.761	-0.033	0.865	-0.061
	mid	0.001	-0.005	-0.012	0.016	0.071	0.715	0.053	0.876	0.059
	mcar	0.013	-0.013	-0.016	0.017	0.001	0.955	0.003	0.938	0.027
	left	0.009	-0.010	-0.011	0.014	0.003	0.941	0.004	0.949	-0.006
PMM	right	0.020	-0.015	-0.017	0.014	-0.016	0.911	-0.011	0.923	0.037
	tail	0.014	-0.012	-0.015	0.015	-0.006	0.928	-0.004	0.933	0.024
	mid	0.014	-0.014	-0.016	0.017	0.005	0.944	0.005	0.947	0.020
	mcar	0.011	-0.009	-0.011	0.010	-0.003	0.955	-0.001	0.948	-0.014
	left	0.012	-0.007	-0.010	0.006	-0.014	0.939	-0.012	0.949	-0.012
2-Part	right	0.013	-0.009	-0.011	0.008	-0.020	0.891	-0.015	0.916	-0.032
	tail	0.014	-0.010	-0.009	0.007	-0.020	0.894	-0.012	0.925	-0.021
	mid	0.007	-0.007	-0.012	0.013	0.014	0.947	0.011	0.949	-0.021
	mcar	0.007	-0.005	-0.008	0.007	0.012	0.954	0.015	0.936	-0.097
	left	0.004	-0.006	-0.009	0.012	0.027	0.925	0.031	0.911	-0.072
MI	right	0.017	-0.006	-0.003	-0.007	-0.038	0.827	-0.027	0.904	-0.131
	tail	0.012	-0.005	-0.005	-0.001	-0.017	0.912	-0.008	0.954	-0.119
	mid	0.002	-0.004	-0.011	0.013	0.034	0.917	0.031	0.913	-0.083
	mcar	0.148	-0.088	-0.115	0.056	0.016	0.591	-0.008	0.504	0.304
	left	0.177	-0.052	-0.067	-0.057	-0.073	0.112	-0.108	0.018	0.145
IRMI	right	0.054	-0.046	-0.109	0.104	0.042	0.278	0.002	0.200	0.225
	tail	0.136	-0.091	-0.108	0.065	0.001	0.708	-0.007	0.571	0.304
	mid	0.115	-0.046	-0.096	0.028	0.023	0.333	-0.034	0.294	0.213
	mcar	0.016	-0.003	-0.010	-0.001	-0.014	1.000	-0.021	1.000	-0.185
	left	0.008	-0.006	-0.009	0.009	-0.012	1.000	-0.009	1.000	-0.174
BGLoM	right	0.004	0.011	0.017	-0.031	-0.074	0.960	-0.052	1.000	-0.204
	tail	0.003	0.007	0.004	-0.013	-0.050	1.000	-0.038	1.000	-0.201
	mid	0.014	-0.006	-0.005	-0.002	-0.004	1.000	-0.004	1.000	-0.176

PMM performed well, as biases of the means of Y_1 and Y_2 are low, their coverage rates are acceptable and plausible and the correlation between Y_1 and Y_2 is accurately retrieved. The two-part model and `mi` also perform quite well, but coverages are much lower for missingness mechanisms that involve the right tail. Also, `mi` yields large correlation biases when the missingness involves the right tail.

`irmi` performance is weak, for all estimates except the bias of the mean. This underperformance of `irmi` is mainly due to the logistic step assigning all missing values to either the point mass or the continuous distribution (c.f. Sections 2.4.4 and 2.4.5).

The BGLoM performs well for all measures, except for the correlation between Y_1 and Y_2. Also, biases for Y_1 and Y_2 are quite large in situations where the missingness mechanism involves the right tail. Maybe coverage of the mean of Y_1 and Y_2 is a bit too well, as coverage rates tend to be 1. Again, it is clear that the BGLoM benefits from the multivariate nature of the data.

Complete case analysis, as expected, shows good results for MCAR, but yields bias in the cross tabulated proportions, low coverages and large mean biases, especially when the left or right tails are involved.

2.6.3 Multivariate skewed with outliers

For the multivariate simulation with outliers, we assessed method performance by comparing the imputed data with the population data. The imputed data depend on the outliers, whereas the population data are considered before the outliers are added.

Log-transforming the data before imputation resulted in a minor improvement for PMM, and the two-part model, but yielded worse results for `irmi`. For `irmi` using robust regression without log-transformation yielded the best results. Given these increases in performance, we present log-transformed results for PMM and the two-part model and 'robust' results for `irmi`. Please note that `mi` always log-transforms semicontinuous data. The results of the multivariate simulation with skewed target variables with outliers can be found in Table 2.5.

It becomes apparent that `irmi` facilitates robust estimation as mean values are very accurately estimated for all missingness mechanisms, except tailed missingness. The two-part model, `mi` and PMM all show larger mean biases, leading to severely lowered coverage rates. We must note that simulation conditions for `irmi` in the case of left-tailed, right-tailed and mid MAR missingness are different from the simulation conditions of the other methods due to algorithmic difficulties with packages that `irmi` depends on. As a solution, we present `irmi` results for these missingness mechanisms with only 25% missingness. Mean biases of the other methods are very similar to those of `irmi` when 25% missingness is imposed.

Curiously, although `irmi` does often yield very accurate imputed means, the coverage rates are always below acceptable levels, indicating that `irmi` does not add enough between variation when considered as a multiple imputation approach.

All investigated methods retrieve the correct proportions for cells A, B, C and D, except for `irmi`. Especially in the case of left and tailed missingness the amount of zeros is wrongly estimated. In the case where the missingness involves the right tail, biases are generally low and coverage rates are acceptable for all methods, except for `irmi`. The performance of `irmi` is rather weak when the right tail is involved.

It is clear that the BGLoM benefits from the multivariate nature of the data. The BGLoM yields acceptable results, although mean biases are sometimes a bit large. Also, the BGLoM yields biased estimates for the correlation when the missingness involves the right tail. Again, BGLoM coverage rates are too large, indicating too much variation between the imputed datasets.

Table 2.5. Outlier simulation: Biases and coverage rates for the mean of the multivariate skewed simulation with outliers. All biases depict the average simulation value (with outliers) subtracted by the population value (without outliers). Please note that the bias in A, B, C and D are observed proportions minus true proportions.

		A	B	C	D	$\bar{Y}_1$	cov$\bar{Y}_1$	$\bar{Y}_2$	cov$\bar{Y}_2$	ρ_{Y_1,Y_2}
CCA	mcar	-0.001	-0.004	-0.003	0.008	0.065	0.956	0.090	0.950	0.357
	left	-0.077	0.008	-0.028	0.098	0.424	0.008	0.482	0.062	0.525
	right	0.068	-0.018	0.013	-0.063	-0.157	0.002	-0.140	0.018	-0.055
	tail	-0.004	0.005	0.007	-0.007	-0.037	0.761	-0.026	0.865	-0.059
	mid	-0.004	-0.014	-0.020	0.039	0.304	0.716	0.377	0.876	0.579
PMM (log)	mcar	0.001	-0.008	-0.007	0.014	0.064	0.756	0.087	0.582	0.336
	left	-0.005	-0.008	-0.005	0.017	0.070	0.608	0.101	0.381	0.373
	right	0.012	-0.012	0.002	-0.002	-0.009	0.929	0.007	0.947	-0.079
	tail	0.002	-0.005	-0.006	0.009	0.013	0.963	0.020	0.928	-0.032
	mid	-0.000	-0.007	-0.008	0.015	0.071	0.678	0.093	0.475	0.376
2-Part (log)	mcar	0.005	-0.007	-0.010	0.014	0.087	0.773	0.105	0.664	0.231
	left	0.010	-0.006	-0.013	0.011	0.071	0.624	0.088	0.582	0.338
	right	0.005	-0.006	-0.008	0.011	0.064	0.948	0.061	0.925	0.068
	tail	0.006	-0.006	-0.008	0.010	0.072	0.940	0.066	0.917	0.116
	mid	0.003	-0.007	-0.012	0.018	0.085	0.571	0.115	0.411	0.314
MI	mcar	0.005	-0.010	-0.010	0.015	0.073	0.736	0.090	0.684	0.086
	left	0.010	-0.012	-0.016	0.019	0.091	0.575	0.116	0.485	0.287
	right	0.013	-0.009	-0.003	-0.001	-0.023	0.887	-0.014	0.923	-0.118
	tail	0.009	-0.009	-0.006	0.006	-0.001	0.941	0.007	0.950	-0.096
	mid	0.000	-0.011	-0.013	0.024	0.119	0.457	0.153	0.344	0.243
IRMI (robust)	mcar	0.166	-0.099	-0.131	0.064	-0.020	0.530	-0.014	0.514	0.444
	left*	0.105	-0.029	-0.044	-0.032	0.018	0.887	0.017	0.891	0.420
	right*	-0.013	-0.036	-0.035	0.083	0.004	0.744	0.027	0.357	0.136
	tail	0.132	-0.081	-0.101	0.049	-0.124	0.043	-0.146	0.079	-0.119
	mid*	0.071	-0.051	-0.056	0.036	0.045	0.658	0.062	0.543	0.472
BGLoM	mcar	0.006	0.007	0.013	-0.026	0.056	1.000	0.037	1.000	-0.026
	left	0.004	0.004	0.002	-0.010	-0.004	1.000	0.024	1.000	0.007
	right	0.020	0.003	0.004	-0.027	-0.083	1.000	-0.043	1.000	-0.201
	tail	0.006	-0.000	0.010	-0.017	-0.040	1.000	-0.061	1.000	-0.201
	mid	0.016	0.009	0.011	-0.035	0.044	1.000	0.061	0.999	-0.005

*IRMI left-tailed, right-tailed and mid MAR missingness were simulated with 25% missingness because of algorithmic problems with large amounts of missingness for continuous parts with outliers.

The BGLoM delivers the most accurate estimate for the correlation between Y_1 and Y_2 when the right tail is not involved. When the right tail is involved, PMM delivers on average the more accurate estimates for the correlation coefficient, especially for tailed MAR missingness.

All in all, there is no one single imputation method for semicontinuous data that is robust against outliers and yields acceptable inference on all investigated estimates across all simulation conditions.

Table 2.6. Comparison between true and imputed ITSR for all imputation methods. Depicted are the total amount of zeros, the amount of values in cells A, B, C and D, the correlation ρ_D of values in cell D, the total correlation ρ, mean ITSR after imputation and the width of the confidence interval.

	zero	A	B	C	D	ρ_D	ρ	mean	ciw
ITSR	69.0	40.00	29.00	0.00	249.00	0.31	0.46	0.07	-
PMM	69.4	36.00	29.40	4.00	248.60	0.31	0.46	0.07	0.02
2-Part	70.4	32.60	30.40	7.40	247.60	0.37	0.47	0.06	0.04
MI	65.33	27.33	25.33	12.67	252.67	0.29	0.36	0.07	0.02
IRMI	55.0	40.00	15.00	0.00	263.00	0.40	0.50	0.07	0.02
BGLoM	92.2	36.60	52.20	3.40	225.80	-0.02	0.01	0.10	0.67

2.7 Application to real data

Two datasets are used for evaluating PMM imputation on real-world data, one from social statistics (The Hague Twitter Scene (HTS) data) and one from official statistics (Dutch Wholesalers Statistics 2008). All investigated variables are either complete or have been edited already. Missingness is imposed by a MAR missingness mechanism.

2.7.1 HTS data

Twitter data gathered from the HTS is chosen as a real-world dataset from social sciences (Sargasso.nl, 2012). Based on the HTS data, Sargasso.nl (2012) created a network indicating the influence of people and their opinions in Dutch politics. The 318 people investigated include politicians, journalists, spin doctors and managers.

One variable that is particularly interesting is the Incrowd Tweet Success Rate (ITSR), indicating for each respondent the percentage of tweets being retweeted or replied within the HTS. This variable is related to the Tweet Success Rate (TSR), being the overall percentage of tweets being replied or retweeted. Both variables are semicontinuous, as some people are never retweeted or replied, but we choose ITSR for demonstration because it contains a larger point mass at zero (22%). Approximately 50% left-tailed MAR missingness was imposed in ITSR with TSR as a covariate.

Table 2.6 shows the results for ITSR after imputation for all investigated methods. PMM estimates the total amount of zeros in the data very accurately. Some values that were originally zeros are set to continuous but overall performance is very good. The same holds for the two-part model, but the two-part model distributes more values into cell C and overestimates the correlation between the continuous parts of cell D. `mi` redistributes values across the four cells, A, B, C and D and underestimates the total amount of zeros. The correlation ρ after imputation is underestimated.

The BGLoM and `irmi` both underestimate the total amount of zeros, although no values that were originally zero are set to continuous. Instead, many values that were originally zero and had a matching continuous value in the covariate TSR are set to continuous after imputation. As a result, the BGLoM underestimates the correlation

Table 2.7. Comparison between true and imputed TEMPS for all imputation methods. Depicted are the total amount of zeros, the correlation between TEMPS and EMPL ρ, mean TEMPS after imputation and the width of the confidence interval.

	zero	ρ	mean	ciw	sum
TEMPS	304.00	0.48	5.02	0.00	4172.00
PMM	312.80	0.50	4.94	2.31	4103.80
2-Part	300.60	0.40	5.88	4.18	4881.98
MI	294.67	0.51	5.02	2.11	4170.49
IRMI	120.00	0.45	5.63	2.70	4681.64
BGLoM	514.00	0.49	4.14	2.26	3440.17

coefficients ρ_D and ρ and `irmi` overestimates these coefficients. The BGLoM severely overestimates the mean of ITSR after imputation.

2.7.2 Dutch Wholesaler Statistics 2008

The Dutch Wholesalers data from 2008 is chosen as a typical real-world dataset from official statistics. The data (N=831 after editing) are collected by Statistics Netherlands (CBS) and consists of variables such as the number of employees, turnover and costs for Dutch wholesalers. We focus on the amount of temporary workers (TEMPS), as this variable has a large point mass at zero (36.5%) and consists otherwise of data that can be considered as continuous.

Approximately 50% left-tailed MAR missingness was imposed (cf. Section 2.3.3) on TEMPS with the total amount of employees (EMPL) as a covariate. Left-tailed missingness is more realistic for this type of data and would be encountered in real life, as the larger companies tend to be always observed in official statistics.

Table 2.7 shows the results for the original data and the investigated methods. PMM performs very well overall and shows low biases in estimating the point mass, the correlation and the mean of TEMPS. The total amount of temporary workers (sum) is closely approximated. The two-part model best estimates the size of the point mass, but the correlation is underestimated, and the mean of TEMPS and the sum of TEMPS are overestimated.

`mi` also performs very well, especially in estimating the sum of TEMPS, but has a bit more bias in estimating the point mass. It shows that the continuous nature of the covariate is beneficial to `mi`. `irmi` underestimates the point mass by a large amount and shows an overestimation of the mean and sum of TEMPS, but bias in the correlation is rather low. The BGLoM shows a large overestimation of the point mass and therefor underestimates the mean and sum of TEMPS, but correlation bias is lowest of all investigated methods.

2.8 Conclusions

How does PMM compare to specialized methods like `mi`, `irmi`, the BGLoM and the two-part model for imputing semicontinuous data? All in all, PMM, `mi` and the two-part model generally outperform `irmi` and the BGLoM.

Between PMM, `mi` and the two-part model, we conclude that PMM performance is best overall. The performance of PMM is at least as good as the performance of `mi` and the two-part model, with PMM often outperforming the other methods. PMM preserves data distributions and imputes only non-negative values when the data consist of non-negative values. `mi` can also impute non-negative values, but the log-transformation procedure leads to imputing non-negative values that are far outside the range of observed values, leaving PMM the only investigated method that preserves the original data distribution.

In the multivariate simulations, it shows that none of the imputation procedures are specifically suitable to impute semicontinuous data in the presence of outliers. Depending on the estimate of interest, it might be beneficial to impute large amounts of incomplete skewed data with outliers by different approaches as there is no single imputation approach that yields acceptable inference over all simulation conditions. Improving on more efficient and robust estimation of predicted means could improve the performance of PMM for semicontinuous data with outliers, but exploring such applications are subject to future work.

An important part of semicontinuous data is the size of the point mass and its relation to auxiliary variables. We can see from both the univariate and multivariate simulations that PMM accurately estimates the size of the point mass, independent from the missingness mechanism, and best preserves the correlation in the data when outliers are not considered. The total amount of zeros and the range and location of the continuous values are also accurately estimated by PMM as estimations for the median and mean yield very low bias. Coverage rates for PMM are acceptable and stable, indicating that standard errors are not too firm or too liberal and that uncertainty and variability within and between imputations are well executed.

The strength of PMM as an imputation method for semicontinuous data lies in its hot-deck properties. Imputed values are drawn from the observed data instead of an assumed model for the distribution. The benefit to this approach is that patterns and relations that are present in the data will be preserved in the imputed data under MCAR and MAR mechanisms, since the missingness mechanism in these models is either random, or based on the observed data. For missing outlying values in very skewed data, there may be no close donor values and model-based predictions can sometimes perform better. Finally, PMM as a hot-deck method requires a sufficiently large donor pool in order to yield acceptable inference.

Our results suggest that PMM can be used by data-analysts and applied researchers as an imputation method for semicontinuous data. However, imputing semicontinuous data in general, must be done with care. Skewness, the missingness mechanism, outliers and the size of the point mass are important factors and may influence the imputations. However, the performance of PMM is very stable and the method

was found to yield accurate inferences in the most extreme conditions, even in the case of no predictive power in the dataset.

Using PMM as an imputation method, instead of the other investigated methods may be convenient in practice. The two-part model, `mi`, `irmi` and the BGLoM are model-based approaches, with accompanying assumptions and limitations. Although some of these limitations can be dealt with by using some kind of transformation of the data, PMM does not rely on these assumptions and does not show the same limitations as these methods. Given that PMM is already available in statistical software gives applied researchers the possibility to use PMM as an all-round imputation method that can be used for other types of data.

There are some limitations to this research. First, we limited our research to continuous covariates. In real datasets, nominal or ordinal data may occur. In practice these type of variables may be handled by using dummy variables or data transformations. We see no reason how that could impact the performance. Second, BGLoM coverage rates often exceed the 95% level. This can be attributed to a too large amount of between variation between the multiply imputed datasets. As a result estimations may be correct on an inference level, but increasing between imputation variance yields too wide confidence intervals, leaving the method to be too conservative. Finally, we display results for simulations with 50% missingness in each variable, thereby severely limiting performance in univariate and multivariate data scenarios. In practice less missingness is often encountered, which will benefit performance of all methods.

To conclude, predictive mean matching is at least as good for imputing semicontinuous data as dedicated methods for such data. PMM is very flexible as a method, due to its hot-deck characteristics, and is free of distributional assumptions. Moreover, PMM tends to preserve the distributions in the data, so the imputations remain close to the data. These properties generally appeal to applied researchers.

3

Partitioned Predictive Mean Matching as a Multi-level Imputation Technique

Summary. Large scale assessment data often has a multilevel structure. When dealing with missing values, such structures need to be taken into account to prevent underestimation of the intraclass correlation. We evaluate predictive mean matching (PMM) as a multilevel imputation technique and compare it to other imputation approaches for multi- level data. We propose partitioned predictive mean matching (PPMM) as an extension to the PMM algorithm to divide the big data multi-level problem into manageable parts that can be solved by standard predictive mean matching. We show that PPMM can be a very effective imputation approach for large multilevel datasets and that both PPMM and PMM yield plausible inference for continuous, ordered categorical, or even dichotomous multilevel data. We conclude that both the performance of PMM and PPMM is often comparable to dedicated methods for multilevel data.

3.1 Introduction

In large scale assessment surveys, missing values for student demographic and socioeconomic background data are frequently encountered. Often such data has a multilevel structure, where respondents are nested within naturally occurring clusters, such as schools or municipalities. Accounting for missingness in data with a multilevel structure is a relatively recent development, and much remains unknown.

In the multilevel analysis model, cluster effects are assumed to be random. Simply ignoring the effects of the cluster during imputation can lead to an underestimation of the intraclass correlation (ICC) in the completed data. Ideally, model parameters within the clusters are allowed to randomly vary during imputation. Although such

This chapter is accepted for publication in Psychological Test and Assessment Modeling as Vink, G., Lazendic, G., & Van Buuren, S. (in press). Partitioned predictive mean matching as a multi-level imputation technique. *Psychological Test and Assessment Modeling*

strategies are available (Zhao and Schafer, 2013), they rely heavily on model assumptions.

Some authors have indicated that multilevel data can also be reasonably imputed when cluster membership is taken into account as a fixed effect (Andridge, 2011; Graham, 2012). A straightforward procedure to include cluster membership into the imputation model as a fixed effect is to use dummy coding strategies. Such strategies should allow for proper estimation of the intraclass correlation, especially when the ICC gets large (Andridge, 2011; Graham, 2012).

Imputing multilevel data by incorporating a fixed effects approach in the imputation model can be very convenient in practice. In such situations, a straightforward extension of current imputation methodology to the situation of multilevel data suffices. For real life data, which does not necessarily follow a specific distribution, such an extension can be especially flexible when an approach is used that does not pose strict assumptions on the distribution of the data.

One imputation approach that is particularly proven to work well in a wide range of situations is predictive mean matching (PMM, Rubin, 1986; Little, 1988). It has been shown that the performance of imputation procedures involving PMM can be very good (Van Buuren, 2012; De Waal et al., 2011; White et al., 2011; Su et al., 2011; Van Buuren and Groothuis-Oudshoorn, 2011; Siddique and Belin, 2007; Yu et al., 2007), especially when normality assumptions are breached or when semicontinuous data is considered (Van Buuren, 2012; Vink et al., 2014). More specifically, PMM does not only yield acceptable and plausible estimates, but also manages to maintain underlying distributions of the data (Van Buuren, 2012; White et al., 2011; Yu et al., 2007; Heeringa et al., 2002; Vink et al., 2014). Implementation of PMM is straightforward in multivariate data problems when the fully conditional specification (FCS, Van Buuren et al., 2006) framework is used. Little is known, however, about the practical applicability of PMM on multilevel data.

We investigate how suitable PMM is as an imputation approach for multilevel data when the clustering of the data is modeled as a fixed effect during imputation. We thereby concentrate on a comparison between PMM, bayesian normal linear imputation (NORM) and a mixture of suitable, dedicated 2-level imputation methods (MIX) in a simulation study. We propose partitioned predictive mean matching (PPMM) as an extension to the predictive mean matching algorithm to facilitate imputation of multilevel data for big datasets. Finally, we apply PPMM to a large real dataset from the Australian Curriculum, Assessment and Reporting Authority (ACARA) to obtain a Bayesian estimate of Social-Educational Advantage on the school level.

3.2 Predictive Mean Matching as a Multilevel Imputation approach

We define $Y = (Y_{obs}, Y_{mis})$ as an incomplete variable, where Y_{obs} and Y_{mis} denote the observed values and the missing values in Y, respectively. We define $X = (X_1, ..., X_j)$

as a set of j fully observed covariates, with X_{obs} and X_{mis} corresponding to the observed and missing parts in Y. Further, n denotes the number of units in Y, n_{mis} and n_{obs} denote the number of units with missing and observed values of Y, respectively, and m denotes the number of multiply imputed datasets to be obtained, with $m \geq 2$. Finally, we require the variable that contains the class structure to be included in X in dummy coded form.

3.2.1 PMM algorithm

Multiply imputing Y_{mis} by means of predictive mean matching can be done by the following algorithm.

1. Use linear regression of Y_{obs} given X_{obs} to estimate $\hat{\beta}$, $\hat{\sigma}$ and $\hat{\varepsilon}$ by means of ordinary least squares.
2. Draw σ^{2*} as $\sigma^{2*} = \hat{\varepsilon}^T\hat{\varepsilon}/A$, where A is a χ^2 variate with $n_{obs}-j$ degrees of freedom.
3. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}(X_{obs}^T X_{obs})^{-1}$.
4. Calculate $\hat{Y}_{obs} = X_{obs}\ \hat{\beta}$ and $\hat{Y}_{mis} = X_{mis}\ \beta^*$.
5. For each missing value $\hat{Y}_{mis,i}$ where $i = 1, \ldots, n_{mis}$:
   a) find $\Delta = |\hat{Y}_{obs,k} - \hat{Y}_{mis,i}|$ for all k, with $k = 1, \ldots, n_{obs}$.
   b) Randomly sample one value from $(\Delta^{(1)}, \ldots, \Delta^{(5)})$, where $\Delta^{(1)}$ through $\Delta^{(5)}$ are the five smallest elements in Δ, respectively, and take the corresponding Y_{obs} as the imputation.
6. Repeat Steps 1 through 5 m times, each time saving the completed dataset.

In the case of multivariate missingness, FCS can be used to iteratively impute every missing datum in each variable of interest, based on a set of covariates. We must note that alternative implementations of PMM do exist (see e.g. (Koller-Meinfelder, 2009; Morris et al., 2014; Schenker and Taylor, 1996; Siddique and Belin, 2007)).

3.2.2 Selecting donors

When performing PMM on multilevel data, three possible scenarios can be used to sample a probable donor value. First, if we ignore the cluster structure, any value in Y_{obs} can in theory be sampled as a donor value, although some values are more likely than others. Assumed that units within a cluster are more alike than units between clusters, this scenario will ignore valuable information, potentially leading to biased results and underestimated cluster effects.

Alternatively, missing values in Y can be imputed by sampling a suitable donor candidate from within the respective cluster. Although potentially very effective, this scenario will quickly pose donor selection problems when clusters size becomes too small or when clusters are completely unobserved.

We prefer a compromise between the first two scenarios. In the algorithm from 3.2.1, the cluster structure is included in the prediction models for $\hat{Y}_{mis}$ and $\hat{Y}_{obs}$. As

a result, the likelihood of a sampled donor value coming from the same cluster (or a similar cluster, for that matter) as the missing value is increased. In this way, the cluster structure is preserved as far as possible, while still allowing for probable donor selection in the case of very small or completely unobserved clusters.

3.2.3 Partitioned predictive mean matching (PPMM)

As datasets become increasingly larger, using dummy coding strategies can become computationally challenging. In the case of many respondents (say 3 million) combined with a large number of clusters (say 10 thousand), expanding the cluster structure to dummy variables may currently even be computationally unfeasible. To avoid computational problems when using predictive mean matching with large multilevel datasets, we propose the following extension to the predictive mean matching algorithm:

1. Partition the data into P approximately equally sized smaller parts, where each part $p = 1, \ldots, P$ contains only whole clusters.
2. Carry out the PMM algorithm from Section 3.2.1 on each part p.
3. Append the P parts for each multiply imputed dataset.

Without loss of generality, the combined data for the P imputed parts can be analyzed conform to current imputation methodology. For the estimation process it is critical that clusters are wholly contained in a single part and are not split among parts. If the data is ordered based on a set of (observed) covariates, such that the likelihood of similar clusters being in the same part is increased, selecting a probable donor can be done on the level of the available donors, without the need for the data as a whole. For example, demographic information can be used to group similar clusters into the same parts. Such a procedure would benefit the imputation model, especially when donor candidates need to be sampled from outside the 'own' cluster.

3.2.4 Speeding up donor selection

Both PMM and PPMM draw imputations from observed values by comparing the distance between each $\hat{Y}_{mis}$ with all $\hat{Y}_{obs}$. This process can become a very lengthy procedure for very large datasets (e.g. $n > 1,000,000$). Using a sufficiently large randomly selected subsample from $\hat{Y}_{obs}$ to sample donors from is computationally convenient and efficient, especially when the number of cases and the proportion of missingness are both large. We propose the following extension to the donor selection step in the PMM algorithm from Section 3.2.1 for large datasets with large amounts of missingness:

1. Draw a subsample $\hat{Y}^S_{obs}$ of length l randomly from $\hat{Y}_{obs}$, with $l < n_{obs}$.
2. Find for each missing value $\hat{Y}_{mis,i}$ the five smallest donors from $\Delta = |\hat{Y}^S_{obs,k_1} - \hat{Y}_{mis,i}|$, where $k_1 = 1, \ldots, l$.

We must note that it is not necessary to choose $l < n_{obs}$, but doing so may greatly benefit computation time.

3.3 Simulation

Predictive mean matching is an imputation method that is relatively easy to implement with a performance that is often very good. To gain insight in the suitability of PMM and PPMM as a multiple imputation approach for multilevel data we performed the following simulation study.

We use the `popularity2` dataset from Hox (2010), a simulated dataset for 2,000 pupils in 100 classes. The dataset contains two level one outcome measures that consider pupil popularity; an indication of pupil popularity (popular, $\mu = 5.08$) derived by a sociometric procedure and pupil popularity as perceived by the teacher (popteach, $\mu = 5.06$). Both outcome variables are measured on a 10-point scale. The explanatory variables are pupil gender (sex, $\mu = 0.51$), pupil's self-measured extraversion (extrav, $\mu = 5.22$) on a 10-point scale, and the experience of the teacher (texp, $\mu = 14.26$) measured in years. The `popularity` data does not consider the school level.

We induce missing completely at random (MCAR) and missing at random (MAR) missingness in the popularity data based on popularity as perceived by the teacher (popteach). Missing values are assigned by using a random draw from a binomial distribution of the same length as Y and of size 1 following the procedure as described in Vink et al. (2014). In the simulations, 15 %, 25 % and 50 % missingness is induced.

To simulate PPMM we partition the data into 10 parts, where each part contains roughly 200 pupils and classes are not split across parts. The average self-perceived pupil popularity differs greatly across classes, which may result in poor performance of the imputation method when using random partitioning, especially when the amount of missingness is large. To this end, the data are sorted based on the average pupil popularity in each class, such that average pupil popularity within parts is more similar than average pupil popularity between parts.

Data imputations are performed with `mice` (version 2.21, Van Buuren and Groothuis-Oudshoorn, 2011) in `R` (version 3.1.0, R Core Team, 2013), with 5 multiply imputed datasets and 10 iterations for the algorithm to converge. PMM and PPMM (both performed by `mice.impute.fastpmm`) are compared to a distributional approach (normal bayesian linear imputation conform `mice.impute.norm`) and a mix of dedicated multilevel imputation algorithms (MIX), namely `mice.impute.2l.norm` for 'extrav', `mice.impute.2l.pan` (based on PAN by Zhao and Schafer (2013)) for 'popular' and `mice.impute. 2lonly.mean` for 'texp'. We use logistic regression imputation (`mice.impute. logreg`) to impute 'gender' for NORM and MIX, but leave out the results under MIX as they are equivalent to the results under NORM.

Table 3.1. Overview of imputation methods used per variable in the simulation

	LEVEL 1			LEVEL 2
	extrav	sex	popular	texp
PMM	pmm	pmm	pmm	pmm
NORM	norm	logreg	norm	norm
MIX	2l.norm	-	2l.pan	2lonly.mean

Each cluster in the popularity dataset contains at least 16 and at most 26 students, with the mode being 20 students. There is no need to investigate larger cluster sizes, as it is known that bias in the ICC decreases as cluster size gets larger (Andridge, 2011).

3.3.1 Evaluation

We evaluate the imputation approaches on the ability to retrieve the following model components for each variable: the average bias of the group means (fixed effect bias), the average coverage rate of the 95 % confidence interval of the group means and the ICC. The ICC is defined as $\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\epsilon^2)$ where σ_α^2 denotes the between group variance and σ_ϵ^2 denotes the within group variance of the random effects. Conveniently, the ICC contains the information about the random effects variance components, therefor we do not evaluate these variances separately.

The above evaluations are carried out on the variable level instead of the model level for two reasons. First, we would like to preserve data structures. Second, the role variables take in a model might change during the analysis stage. Outcome variables in one model may become predictors in another model, and vice versa. Ultimately, it would be ideal if both models can be analyzed on the same data.

Because we consider the popularity data as the population and induce missingness in the data directly, we have no sampling variation and the pooling rules proposed in Vink and Van Buuren (2014) are used.

3.3.2 Results

Intraclass correlation

The popularity data displays strong population intraclass correlations (see Table 3.2). All imputation methods yield results that are relating close to these population values. As expected, the bias increases with the amount of missingness.

The experience of the teacher (texp) is particularly interesting when considering the intraclass correlation. Because the experience is the same for all pupils in a cluster, the intraclass correlation equals 1. PMM is able to automatically replicate this structure, yielding the correct inference as if the data were deductively imputed. NORM does not sample from the observed data, but rather draws from a normal distribution, resulting in a small deviation from the population value. MIX uses a cluster mean imputation method, which yields unbiased results when at least one pupil is observed in each cluster. For larger amounts of missingness, however, it may occur that clusters are completely unobserved. In such situations, MIX is not able to find an imputation.

The slightly larger bias for the ICC for teacher experience in the case of PPMM can be explained by the correlation between pupil popularity and teacher experience ($\rho(1998) = .29, p < .01$). Sorting the data based on pupil popularity will have an effect on the distribution of teacher experience over parts. Together with the smaller sample size, this results in occasional imputations for teacher experience that are different from the observed values.

Fixed effects

The fixed effects are very accurately estimated by all imputation approaches, see Figure 3.1. PMM displays a very stable performance, with very low variation in the bias across missingness mechanisms. PPMM shows more variation in the case of 50 % missing data, which can be explained by restrictions put on the available donor pool due to partitioning. For all methods it holds that bias is very small, even for large amounts of missingness.

Because teacher experience is a level-2 variable, values are the same for every pupil in a cluster. For such variables, PMM will yield correct results, even when none of the cluster's values are observed. MIX will also yield correct results, but only for clusters that have at least one observed value.

Coverage rates of the cluster means

Figure 3.2 displays the average over the coverage rates for the cluster means. It can be seen that performance for all methods is very stable across all variables, with very good coverage rates under all missingness mechanisms when considering 15 percent

Table 3.2. Bias of the intraclass correlations after imputation as deviations from the population value (truth).

		15 % missing				25 % missing				50 % missing			
meth	mech	extrav	sex	texp	pop	extrav	sex	texp	pop	extrav	sex	texp	pop
TRUTH	-	0.262	0.112	1	0.363	0.262	0.112	1	0.363	0.262	0.112	1	0.363
	mcar	0.009	0.002	0	0.004	0.017	0.009	0	0.008	0.047	0.036	0	0.021
	left	0.007	0.004	0	0.002	0.014	0.010	0	0.004	0.053	0.048	0	0.017
PMM	right	0.011	0.002	0	0.001	0.020	0.008	0	0.006	0.056	0.044	0	0.020
	mid	0.009	0.003	0	0.004	0.017	0.010	0	0.008	0.050	0.046	0	0.025
	tail	0.009	0.002	0	0.000	0.015	0.006	0	0.003	0.050	0.031	0	0.015
	mcar	0.007	0.001	-0.001	0.006	0.012	0.004	-0.001	0.011	0.034	0.027	-0.003	0.030
	left	0.006	0.005	-0.001	0.005	0.016	0.013	-0.001	0.007	0.048	0.047	-0.002	0.029
PPMM	right	0.004	-0.009	-0.001	0.006	0.013	-0.012	-0.002	0.009	0.033	0.011	-0.007	0.030
	mid	0.007	0.004	-0.001	0.004	0.012	0.011	-0.001	0.008	0.035	0.041	-0.002	0.031
	tail	0.005	-0.004	-0.001	0.004	0.010	-0.004	-0.001	0.007	0.034	0.011	-0.003	0.026
	mcar	0.007	-0.009	≈ 0	0.003	0.012	-0.009	≈ 0	0.004	0.043	0.013	≈ 0	0.018
	left	0.006	-0.009	≈ 0	0.001	0.011	-0.004	≈ 0	0.003	0.042	0.036	≈ 0	0.015
NORM	right	0.011	-0.011	≈ 0	0.001	0.020	-0.010	≈ 0	0.005	0.057	0.024	≈ 0	0.022
	mid	0.007	-0.003	≈ 0	0.003	0.014	-0.001	≈ 0	0.006	0.042	0.026	≈ 0	0.018
	tail	0.007	-0.016	≈ 0	-0.000	0.014	-0.017	≈ 0	0.001	0.048	0.007	≈ 0	0.011
	mcar	-0.001	-	0	-0.000	-0.006	-	0	-0.001	-0.018	-	0	0.003
	left	-0.006	-	0	-0.002	-0.012	-	0	-0.003	-0.030	-	*0	0.001
MIX	right	-0.006	-	0	-0.000	-0.011	-	0	0.002	-0.019	-	*0	0.011
	mid	-0.001	-	0	-0.000	-0.005	-	0	-0.000	-0.020	-	0	0.005
	tail	-0.007	-	0	-0.000	-0.012	-	0	0.000	-0.031	-	0	0.003

*Values are calculated based on the imputed clusters only, due to some clusters being completely unobserved.

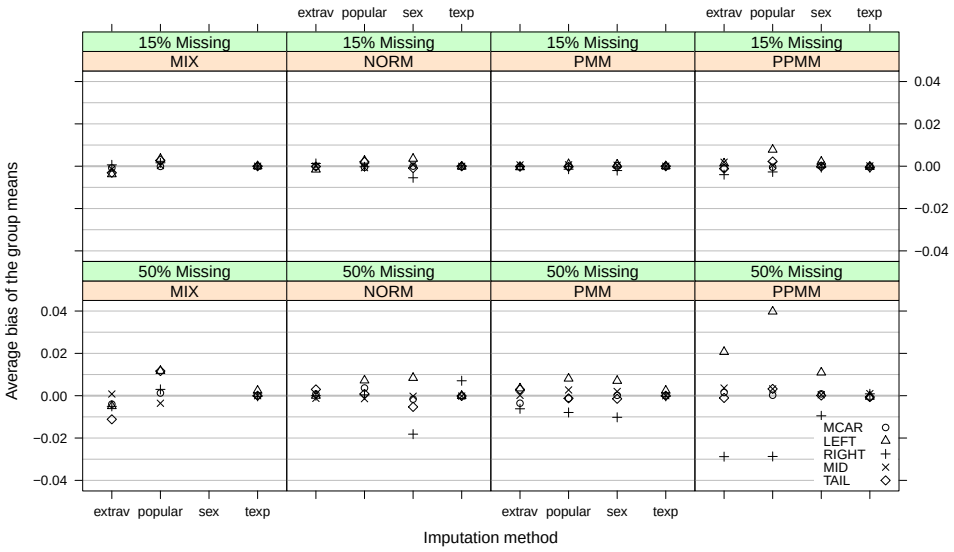

Fig. 3.1. Average bias of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

missingness. Because of the unbiased group means, PMM is able to perfectly cover the average teacher experience in the population after imputation, even for large missingness percentages. When missingness increases to 50 percent, PMM performance is still good for all variables. However, the dichotomous variable gender may be more efficiently imputed using logistic regression imputation (as under NORM).

Confidence interval width

Confidence interval widths are generally small when the cluster structure is taken into account, with PMM yielding slightly smaller intervals than the other imputation approaches. As expected, the average interval width for teacher experience under PMM is zero and the average interval width under NORM is close to zero. For MIX, the average interval width is unbiased, but is calculated over the observed clusters for large amounts of missingness. As expected, interval widths between PMM and PPMM are very similar, with the exception of the interval widths for experience of the teacher.

3.4 Application

We apply PPMM on a dataset collected by ACARA for the purpose of providing fair and meaningful comparisons of student performance in the National Assessment

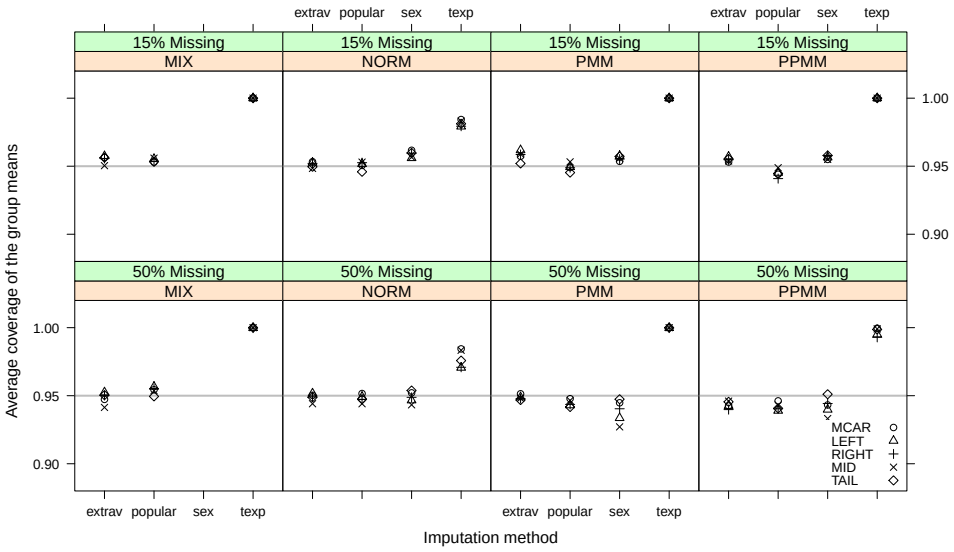

Fig. 3.2. Average coverage rate of the 95 percent confidence interval of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

Program - Literacy and Numeracy (NAPLAN) between schools serving students from statistically similar socio-educational backgrounds. The resulting student background dataset (SBD) contains 2.782.060 Australian students clustered in 9.671 Australian schools and can be used for obtaining an estimate of the social educational advantage (SEA) score for Australian schools. An overview of the most important variables in the dataset is given in Table 3.3.

All variables with missingness are imputed, but here we focus on parent education and occupation. The parental variables are ordered categorical variables, with 'occupation' and 'non-school education level' having a separate category that records 'not in paid work' and 'no non-school qualification', respectively (see Table 3.4). The dual (or semi-categorical) nature of these data can be split in two parts: an ordered distribution over the categories and a point mass that does not follow the ordering of the other categories. For continuous or integer data with such a point mass, PMM is known to be a very effective single-step imputation approach (Vink et al., 2014).

Another reason for focusing on the parent variables is the large amounts of missingness. The parent variables contain most of the missing values in the data, ranging from 17 to 32 percent missingness per variable. As a result, a large number of observations may be missing on the school level and schools are sometimes even completely unobserved. With such large amounts of missingness, it is important to use an im-

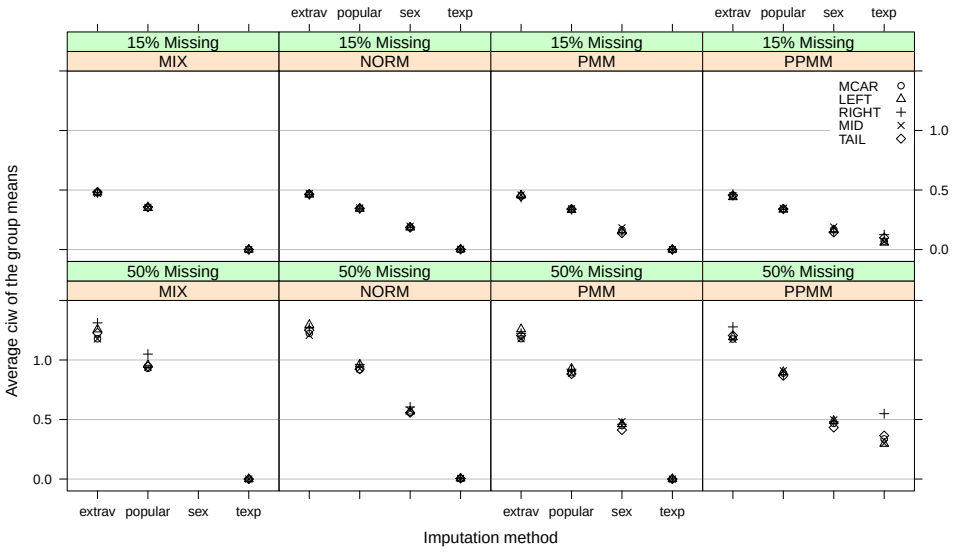

Fig. 3.3. Average width of the 95 percent confidence interval of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

Table 3.3. Variables in the SBD dataset

variable name	level	description	%mis
school_ID	2	school identifier	0 %
jurisdiction	2	school jurisdiction	0 %
sector	2	school sector	0 %
geolocation	2	school geographical location	2.21 %
sex	1	pupil gender	0.37 %
indigenous_status	1	pupil indigenous status	1.72 %
year_level	1	pupil year level	0.76 %
parent1_educ_schl	1	parent 1 school education level	17.30 %
parent1_educ_nonschl	1	parent 1 non-school education level	16.96 %
parent1_occ	1	parent 1 occupation	22.58 %
parent2_educ_schl	1	parent 2 school education level	29.92 %
parent2_educ_nonschl	1	parent 2 non-school education level	29.35 %
parent2_occ	1	parent 2 occupation	31.97 %

putation procedure that is able to capture the multilevel structure. Neglecting the clustering of the data will result in an underestimation of the intraclass correlation.

Finally, the parent variables are critical in the direct estimation of the SEA score for a school. See Acara (2014) for a detailed explanation of the modeling of Australian social educational advantage measures.

3.4.1 Procedure

We partitioned the SBD data into 271 parts based on jurisdiction, sector and geolocation, with parts containing only whole schools and schools not being split among parts. Each partition contains approximately 10,000 cases. Imputations are performed using PPMM with 5 imputations and 10 iterations for the algorithm to converge. We set $l = 1,000$ as the sample size for the random subset of observed donor candidates to speed up the imputation process.

3.4.2 Results

After imputation, the intraclass correlation in the parent variables is similar to the intraclass correlation of the parent variables in the incomplete data (see Table 3.5),

Table 3.4. Levels of the parent variables in the SBD

Parent occupation
Senior management and qualified professionals
Business managers and associate professionals
Tradesmen/women, clerks and skilled staff
Labourers and related workers
Not in paid work in last 12 months
School education level
Year 12 or equivalent
Year 11 or equivalent
Year 10 or equivalent
Year 9 or equivalent or below
Non-school education level
Bachelor degree or above
Advanced diploma/Diploma
Certificate I to IV (including trade certificate)
No non-school qualification

Table 3.5. Intraclass correlations and average group means in the observed and imputed data. Shown are the average imputation value ($\hat{m}$) and the observed (but incomplete) data estimate (obs) for education (schooled and non-schooled) and occupation for both parents.

	ICC		MEANS	
	$\hat{m}$	obs	$\hat{m}$	obs
parent1_educ_schl	0.19	0.17	3.17	3.17
parent1_educ_nonschl	0.05	0.05	6.77	6.76
parent1_occ	0.18	0.17	4.36	4.32
parent2_educ_schl	0.20	0.18	3.02	3.04
parent2_educ_nonschl	0.05	0.06	6.56	6.55
parent2_occ	0.24	0.22	3.20	3.16

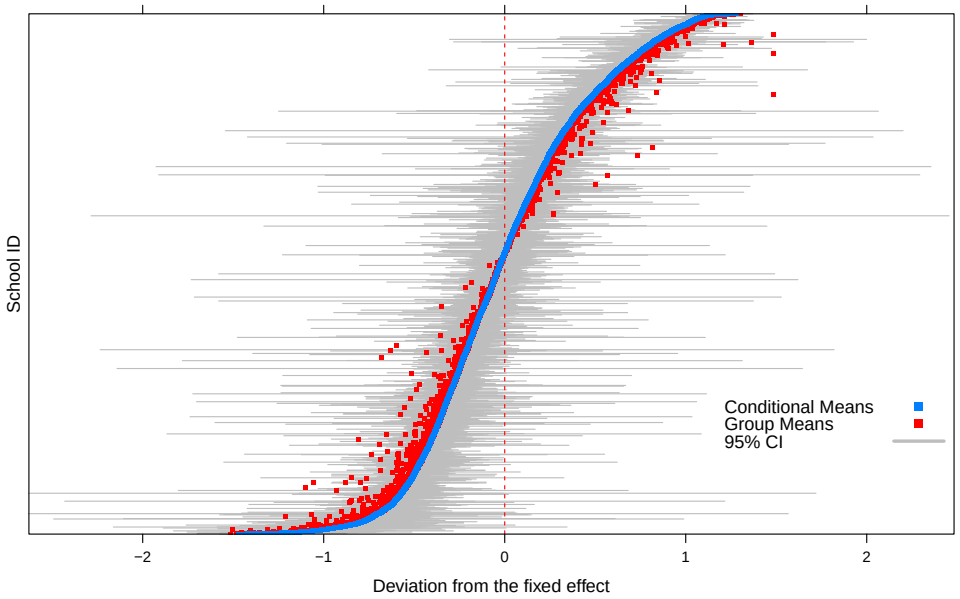

Fig. 3.4. Conditional SEA means from the random effects model and group SEA means from the fixed effects model compared after imputation. Shown are pooled results for the conditional means, the group means and the 95 % confidence interval for the conditional means.

with difference being generally very small. This indicates that the utilized imputation method was able to give meaningful predictions, thereby taking group membership into account.

Table 3.5 also shows the fixed effect estimates of the parent variables before and after imputation. It can be seen that difference is very small, when compared to the observed values, indicating that PPMM is able to very accurately estimate fixed effect from the incomplete data.

We used a random effects model to estimate social educational advantage in each of the imputed datasets. The model takes the form $SEA_{ab} = \mu + U_a + W_{ab}$ where SEA_{ab} is the score of the bth pupil at the ath school, μ is the overall average, U_a is the school-specific random effect and W_{ab} is the individual-specific error. The estimates and variances from the five imputed datasets were combined to obtain a Bayesian estimate for social educational advantage on the school level.

In Figure 3.4, we compare the conditional means from the random effects model to the group means from the fixed effects model. Note that the data has been sorted based on size of the conditional means. It can be clearly seen that the random effects

model takes the group size into account and that shrinkage towards the fixed effect is applied.

The larger confidence interval widths belong to the smaller schools that are completely unobserved. As a result of the increased uncertainty caused by the large amounts of missingness in these schools, the between imputation variance and, naturally, the confidence interval width increases.

3.5 Discussion

PMM emerges as a very effective imputation technique for multilevel data when the cluster structure is taken into account. The algorithm is able to preserve the multilevel nature of the data, leading to precise and well-covered estimates.

Controlling for cluster membership is not the only requirement for a good multilevel imputation approach. In our view an imputation method for multilevel data must adhere to the following properties:

1. *Structure preserving:* The cluster structure should be accounted for during imputation.
2. *Generality:* The cluster size and the amount of missingness may vary and clusters may be completely unobserved.
3. *Observed plausibility:* Imputed values must be within the range of plausible values such that only realistic values can be imputed.

In a multilevel setting it can be concluded that PMM performance is comparable to dedicated methods for multilevel data and that PMM is sometimes even able to outperform dedicated methods, especially when the amount of missingness is large or when some clusters are completely unobserved.

For small cluster sizes and 50 percent missing data, there can be a slight (but conservative) overestimation of the ICC. However, in all simulation conditions PMM yields realistic imputations that are within the bounds of the plausible data values. In practice, this proves to be especially convenient when imputing continuous ratio scales, dichotomous variables, categorical variables, or even semicontinuous data.

We proposed partitioned predictive mean matching as a straightforward extension to the PMM algorithm that divides the big-data multilevel problem into manageable parts that can be solved by standard predictive mean matching. We have demonstrated that PPMM performance is similar to the performance of unpartitioned PMM, proving PPMM to be an effective imputation approach for large datasets, especially those datasets where dummy coding strategies are computationally not feasible.

The continuous variables in the simulation study in Section 3.3 are normally distributed. It is well known that deviations from normality can have a serious impact on the performance of methods that assume such distributions (MIX, NORM): evaluating the performance of these methods on non-normal data would be pointless. PMM, on the other hand, is known to handle deviations from normality very well (Vink

et al., 2014). Our application on real data also demonstrates that PMM can still be used in situations where non-normal distributions are considered.

PMM is widely recognized as a method that preserves data distributions and, although it uses underlying methodology that assumes variables to be continuous, we have shown that it can yield plausible inference for continuous, ordered categorical, or even dichotomous multilevel data.

Part II

Bivariate imputation

4

Multiple Imputation of Squared Terms

Summary. We propose a new multiple imputation technique for imputing squares. Current methods yield either unbiased regression estimates or preserve data relations. No method, however, seems to deliver both, which limits researchers in the implementation of regression analysis in the presence of missing data. Besides, current methods only work under a missing completely at random (MCAR) mechanism. Our method for imputing squares uses a polynomial combination. The proposed method yields both unbiased regression estimates, while preserving the quadratic relations in the data for both missing at random and MCAR mechanisms.

4.1 Introduction

Multiple imputation (MI) is the method of choice for many incomplete data problems. MI incorporates the uncertainty about the missing data by creating $m > 2$ imputed data sets. Missing values are filled in under an *imputation model*. The imputed data that result from the imputation model is then analyzed by the *analysis model*. Separate analyses can be combined to get a single inference or set of estimates by making use of the combining rules derived by Rubin (1987).

The most critical part of MI is specification of the imputation model. It is widely accepted that the imputation model should embrace all relations of scientific interest. Usually, this is done by incorporating the variables of interest as main factors. However, things become less clear if the scientific model contains nonlinear terms.

As an example, if we want to predict Y from X and its square X^2, then both X and X^2 should be included in the imputation model. Leaving the term X^2 out of the imputation model will result in a downward bias of the slopes when we perform a regression analysis on the imputed data. However, although it is generally agreed

This chapter is published as Vink, G., & Van Buuren, S. (2013). Multiple Imputation of Squared Terms. *Sociological Methods & Research*, *42*(4), 598-607.

that all squares and interactions should be accounted for in MI, no consensus on how to do this has been reached.

Von Hippel (2009) reviewed several approaches to imputing squares. The *'transform, then impute'* method calculates the squares and interactions in the incomplete data for the cases that have no missing values, and then imputes the derived variable like any other variable. The *'impute, then transform'* method imputes variables in their raw form, and then calculates the derived variable in the imputed data after imputation. These methods were compared to the *passive imputation* method (Van Buuren et al., 1999), implemented in the `mice` package in R (Van Buuren and Groothuis-Oudshoorn, 2011), and the *ice* command for Stata (Royston, 2005).

Von Hippel (2009) advises to use the transform-then-impute method, which delivers acceptable regression estimates but heavily distorts the relationship between X and X^2. Figure 4.1 shows that for the transform-then-impute method, imputations do not follow the relation in the population (observed) data. We agree with Von Hippels conclusion, but do not want to overlook that the transform-then-impute method yields combinations of imputed values that would never occur, had the data been observed. Such imputations are implausible and should be rejected on that ground.

We must note that Von Hippel's conclusions are based on a missing completely at random (MCAR) mechanism (Seaman et al., 2012), where the missingness does not depend on the data, which is a limitation in practice. An imputation method would be more powerful if it yields acceptable inference under the missing at random (MAR) mechanism, where the missingness may depend on the data, but must not depend on the missing data itself.

Because existing methods for imputing squared terms are severely limited, we propose the *polynomial combination* approach, which yields unbiased regression estimates, while at the same time preserving the consistency between the imputed values, for MAR and MCAR mechanisms.

4.2 Method

4.2.1 Formulation of the problem

The model of scientific interest is

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon \tag{4.1}$$

with $\epsilon \sim N(0, \sigma^2)$. We assume that Y is complete and that $X = (X_{obs}, X_{mis})$ and $X^2 = (X^2_{obs}, X^2_{mis})$ are partially missing. The problem is to find imputations for X such that estimates of α, β_1, β_2, and σ^2 are unbiased, while ensuring that the quadratic relation between X and X^2 will also hold in the imputed data.

4.2.2 Polynomial combination method

Define the polynomial combination $Z = (Z_{obs}, Z_{mis})$ as the linear combination $Z = X\beta_1 + X^2\beta_2$. The idea is to impute the missing values in Z instead of X and X^2,

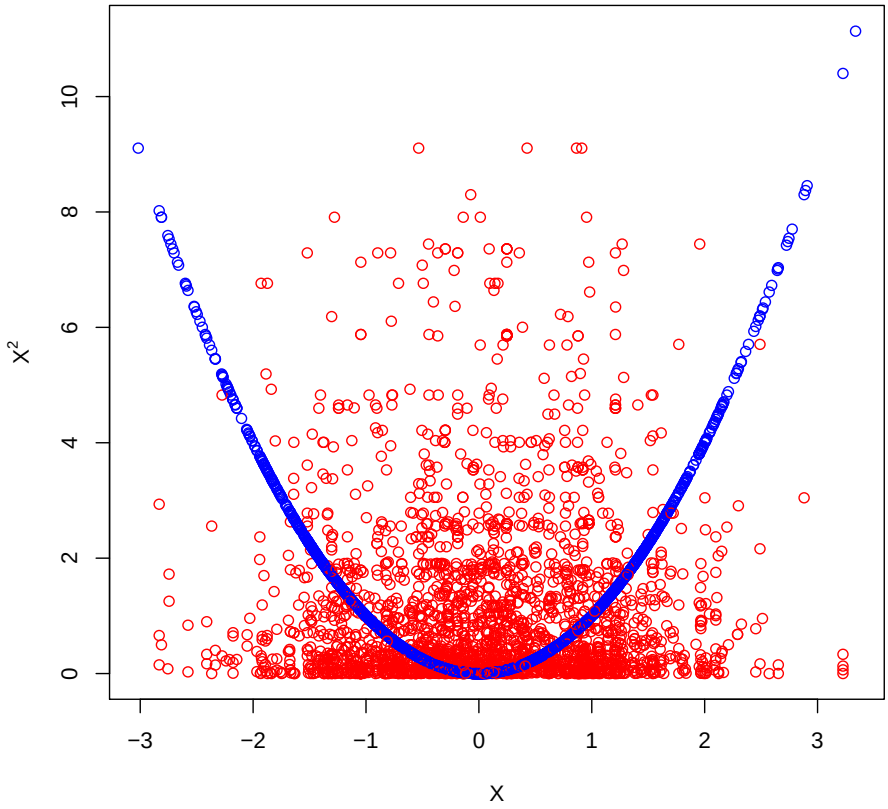

Fig. 4.1. Transform-then-impute imputations. Observed (blue) and imputed values (red) for X and X^2.

followed by decomposing the imputed data Z into components X and X^2. Imputing Z reduces the multivariate imputation problem to a univariate problem, which is easier to manage.

Under the assumption that $P(Y, Z)$ is multivariate normal, we can impute the missing part of Z as $Y\beta^* + \epsilon^*$. Here β^* is a random draw from the posterior distribution of the linear regression of Y on Z, and ϵ^* is a draw from the residual distribution $Z - Y\hat{\beta}$. In cases where the normal residual distribution is unrealistic, we can use predictive mean matching (PMM) Little (1988).

The next step is to decompose Z into X and X^2. Under Model (4.1) this is straightforward. The imputed value Z has two distinct real roots:

$$X_- = -\frac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} + \beta_1\right) \tag{4.2}$$

$$X_+ = \frac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} - \beta_1\right) \tag{4.3}$$

where the discriminant $\Delta = 4\beta_2 Z + \beta_1^2$ must be greater than zero. The case $\Delta = 0$ occurs if and only if both β_1 and β_2 are exactly 0, resulting in just one distinct real root, namely $X_0 = -\beta_1/2\beta_2$. Since incorporating nonexistent relationships in the analysis serves no further purpose, we assume that regression estimates are always unequal to 0.

Given this assumption, for any given Z, we can take either $X{=}X_-$ or $X{=}X_+$, and square it to obtain X^2. Either root is consistent with $Z = X\beta_1 + X^2\beta_2$, but choice among these two options requires care. Note that the minimum of the parabola is located at $X_{min} = -\beta_1/2\beta_2$. If we choose X_- for all Z, then all imputed $X \leq X_{min}$ will correspond to points located on the left arm of the parabolic function. This is generally not as intended. A sampling mechanism to determine whether to choose from X_- or X_+ for a given Z is needed.

The choice between the roots is made by random sampling, conditional on Y, Z and their interaction YZ. Let $V = (V_{obs}, V_{mis})$, where V_{obs} is a binary random variable defined as 0 if $X_{obs} \leq X_{min}$ and 1 otherwise. We model the probability $P(V = 1)$ by logistic regression as

$$\text{logit}P(V = 1) = Y\beta_Y + Z\beta_Z + YZ\beta_{YZ} \tag{4.4}$$

on the observed data. Assuming that the same model applies to the missing values in X (i.e., that the missingness mechanism is ignorable), we calculate the predicted probability $P(V = 1)$. As a final step, a random draw from the binomial distribution is made, and the corresponding (negative or positive) root is selected as the imputation. This is repeated for each missing value.

4.2.3 Imputation algorithm

The procedure leads to the following algorithm for imputing squares:

1. Calculate X^2_{obs} for the observed X
2. Use PMM to multiply impute X_{mis} and X^2_{mis} as if they were unrelated, resulting in imputations X^* and X^{*2}.
3. Estimate the pooled estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ by linear regression of Y, given $X = (X_{obs}, X^*)$ and $X^2 = (X^2_{obs}, X^{*2})$
4. Calculate the polynomial combination $Z = X\hat{\beta}_1 + X^2\hat{\beta}_2$
5. Multiply impute Z_{mis} by PMM, resulting in imputations Z^*
6. Calculate roots X_- and X_+ given $\hat{\beta}_1$, $\hat{\beta}_2$ and Z^* using Equations(2) and (3)
7. Calculate the abscissa at the parabolic minimum/maximum $X_{min} = -\hat{\beta}_1/2\hat{\beta}_2$
8. Calculate $V_{obs} = 0$ if $X_{obs} \leq X_{min}$, else $V_{obs} = 1$

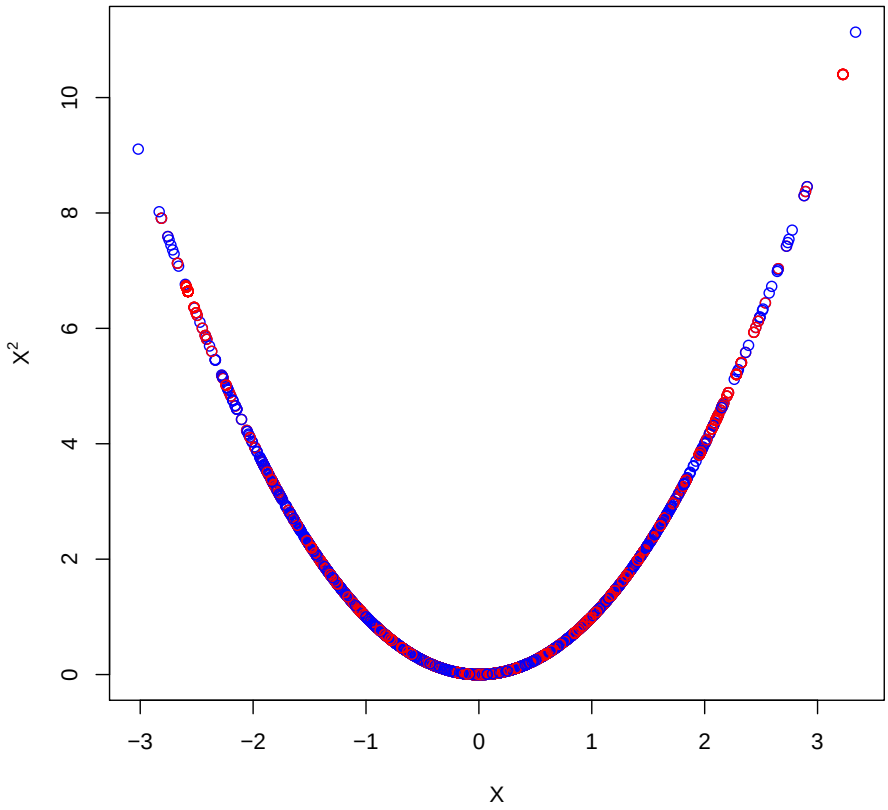

Fig. 4.2. Polynomial combination imputation. Observed (blue) and imputed values (red) for X and X^2.

9. Impute V_{mis} by logistic regression of V given Y, Z and YZ, resulting in imputations V^*
10. If $V^* = 0$ then assign $X^* = X_-$, else set $X^* = X_+$
11. Calculate X^{*2}

The imputations Z^* will satisfy $Z^* = X^*\hat{\beta}_1 + X^{*2}\hat{\beta}_2$.

4.3 Results

To illustrate the polynomial combination method, we simulated and compared the performance of all methods discussed by Von Hippel (2009) against the polynomial

combination method. Data were generated according to the model $Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon$, where X is randomly generated from a standard normal distribution. A larger sample size ($n = 10,000$) was chosen to demonstrate convergence. However, the method works well for smaller sample sizes. We fixed the population intercept α at 0 and the residual standard deviation σ_ϵ at 1. Deviations seem to be larger when the slope of both X and X^2 are larger, hence the population slopes β_1 and β_2 were set to 1. Let R be a response indicator with

$$R = \begin{cases} 1 & \text{if } X \text{ is observed} \\ 0 & \text{if } X \text{ is missing} \end{cases}. \tag{4.5}$$

and let Z_{mis} denote the missing values in Z. Given these settings we created 50 percent joint missingness in X and X^2 according to four MAR mechanisms that follow

$$P(R = 0|Z_{obs}, Z_{mis}, Y) = P(R = 0|Z_{obs}, Y), \tag{4.6}$$

by using a random draw from a binomial distribution of the same length as Y and of size 1 with missingness probability equal to the inverse logit

$$P(R = 0) = \frac{e^a}{(1 + e^a)}.$$

Setting $a = (-\bar{X} + X_i)/SD_X$ gives 50 percent left-tailed MAR missingness. Right-tailed, centered and tailed MAR missingness can be created by setting $a = (\bar{X} - X_i)/SD_X$, $a = .75 - [(\bar{X} - X_i)/SD_X]$ and $a = -.75 + [(\bar{X} - X_i)/SD_X]$, respectively. Adding or substracting a constant moves the sigmoid curve, which results in different missingness proportions.

As an analysis, we used linear regression to see whether the population values could be estimated after imputation. We repeated the analyses 100 times.

The regression estimates after applying the polynomial combination imputation can be found in Table 4.1. The estimated coefficients of the imputed X and X^2, the coefficient of the intercept α and the residual standard deviation σ_ϵ are all close to their respective population values. Missingness mechanisms that involve the right tail show slightly larger deviations.

In contrast, Table 4.1 also displays the performance of the impute-then-transform method regression estimates under the same simulation conditions. The impute-then-transform method yields biased regression estimates, even under MCAR.

Table 4.1 also shows the performance of the passive imputation method. Passive imputation performance is similar to the problematic performance of the impute-then-transform method, as both methods calculate X^2 afterwards.

Finally, the transform-then-impute method yields unbiased regression estimates, but only for MCAR. Although some estimates are retrieved, performance is severely impaired under the MAR assumption (see Table 4.1).

All in all, the polynomial combination method yields regression estimates that are both unbiased and preserve the data relation between X and X^2. The polynomial combination method also perfectly reproduces the population relation between X and

	Missingness Mechanism				
	MCAR	MARleft	MARmid	MARtail	MARright
Polynomial combination					
Intercept (α)	0	-0.01	-0.01	-0.05	-0.07
Slope of X (β_1)	1	1	1	0.96	0.96
Slope of X^2 (β_2)	1	1	1.01	1.06	1.09
Residual SD (σ_ϵ)	1	1	1	1.03	1.05
R^2	0.75	0.75	0.75	0.73	0.73
Impute, then transform					
Intercept (α)	0.39	0.29	0.26	0.52	0.56
Slope of X (β_1)	0.93	0.94	0.87	1.01	1.06
Slope of X^2 (β_2)	0.61	0.60	0.67	0.56	0.66
Residual SD (σ_ϵ)	1.48	1.44	1.41	1.56	1.62
R^2	0.45	0.48	0.5	0.39	0.34
Passive imputation					
Intercept (α)	0.39	0.29	0.26	0.52	0.56
Slope of X (β_1)	0.93	0.94	0.87	1.01	1.05
Slope of X^2 (β_2)	0.61	0.60	0.68	0.56	0.66
Residual SD (σ_ϵ)	1.48	1.45	1.41	1.57	1.62
R^2	0.45	0.48	0.50	0.38	0.34
Transform, then impute					
Intercept (α)	0	0.19	-0.13	0.01	-0.05
Slope of X (β_1)	1	0.91	0.97	1.14	1.32
Slope of X^2 (β_2)	1	0.91	0.95	1.14	1.32
Residual SD (σ_ϵ)	1	0.95	1	1.06	1.15
R^2	0.75	0.77	0.75	0.72	0.67

Table 4.1. Average parameter estimates for different imputation methods under five different missingness mechanisms over 100 imputed datasets ($n = 10,000$) with 50% missing data. The population parameters are $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_\epsilon = 1$ and $R^2 = .75$

its square X^2 in the imputed data. See Figure 4.2 for a graphical representation of the population and imputed data relations between X and X^2, as generated by the polynomial combination method.

We also looked at the mean and covariance matrix as reproduced by the imputed data and compared it to the population values. The mean and covariance matrix of (X, X^2, Y) are

$$\mu = \begin{bmatrix} 0 \\ 1 \\ \beta_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & & \\ 0 & 2 & \\ \beta_1 & 2\beta_2 & 1 + \beta_1^2 + 2\beta_2^2 \end{bmatrix} \tag{4.7}$$

A set of k mean values can be pooled to a single residual mean value with

$$\Delta_\mu = \frac{1}{k}\sum_{i=1}^{k}|\mu_i - m_i| \tag{4.8}$$

where m_i is the ith mean value for the imputed data. Likewise, a pooled residual covariance matrix can be created by

$$\Delta_\Sigma = \frac{1}{k}\sum_{i=1}^{k}|\Sigma_i - S_i| \tag{4.9}$$

where S_i is the ith covariance matrix of the imputed data. Performing a small simulation of $n = 100$ with various regression weights and combining the results with Equations (4.8) and (4.9), yields the following pooled residual mean and covariance matrix.

$$\Delta_\mu = \begin{bmatrix} 0.003 \\ -0.004 \\ -0.003 \end{bmatrix} \quad \text{and} \quad \Delta_\Sigma = \begin{bmatrix} -0.004 & & \\ 0 & 0.007 & \\ -0.004 & 0 & -0.012 \end{bmatrix} \tag{4.10}$$

The results in Equation (4.10) suggest that the mean and covariance matrix in the population data are accurately preserved in the imputed data. Given that only normal imputations that preserve the mean and covariance matrix from the population data can yield unbiased imputations, we can now confidently say that the polynomial combination method yields unbiased regression estimates and delivers transformed variable imputations that are consistent with each other.

All computations in this study have been carried out in R and all imputations are generated with the `mice` package in R (Van Buuren and Groothuis-Oudshoorn, 2011) with $m = 5$ multiple imputations. A `mice.impute.quadratic` routine that implements the polynomial combination method is available in `mice`.

4.4 Conclusion

The polynomial combination method as developed here provides unbiased estimates for problems where incomplete X and X^2 are both in the complete data model. It merges imputation techniques and decomposition of the quadratic equation to obtain the same unbiased regression estimates as the basic transform-then-impute method, while preserving the relations between X and X^2. Also, it performs well under both MCAR and MAR missingness mechanisms. Our advice is to use the polynomial combination method to impute transformed variables with squared relations.

We note that the simulation conditions used are rather harsh. For example, 50 percent of X is missing and some missingness mechanisms severely limit the amount of usable predictive information, especially right-tailed MAR missingness. Also, note that imputations are based on just one covariate. In real-life data sets, conditions for imputing the data are often much better. Yet, also for simpler incomplete data problems, the polynomial combination method yields the best possible inferences even though the difference with the results from other methods may be smaller.

We limited our calculations and analyses to squares, which are essentially interactions between two identical variables. Interactions between different variables remain best imputed using the transform-then-impute method. The polynomial combination method can be generalized to more complex non-linear combinations. We expect that the proposed method also applies to problems in which the scientifically interesting model contains multiple versions or transformations of X, such as interactions between different variables, higher degree polynomial equations and perhaps even splines, which are essentially piecewise polynomials. Exploring such applications of the polynomial combination method is subject to future work.

5

Predictive Ratio Matching Imputation of Nested Compositional Data with Semicontinuous Variables

Summary. Imputing compositional data is challenging because imputations must obey the restrictions in the data while remaining strictly non- negative. The usual methods yield imputations that do obey the re- strictions, but may severely distort the distributions and the relations among the components. We propose predictive ratio matching (PRM) as a general imputation method for compositional data. PRM imputes compositional data by iteratively updating the pairwise ratios in the data. The proposed method emerges as a very effective imputation approach for nested compositional data that can handle the skewed semicontinuous variables. Further, PRM yields imputations that obey sum restrictions, while keeping the data distributions and relations among components intact.

5.1 Introduction

Compositional data can be defined as a set of parts that obey a certain edit restriction, such that the parts have to sum up to a certain total. These parts can be considered as proportions that sum up to 1 or, perhaps more conveniently, as raw data that sums up to a certain total. In both approaches to compositional data, the information is contained in the ratios between each part and the total. Compositional data in the form of raw data is often incompletely encountered in official statistics, such as income components or expenditure of household budget. The missing values in the data pose problems in carrying out and interpreting analyses that assume the data to be completely observed.

Let us consider x_1 as a combination of components $x_2, \ldots, x_D$ such that

$$x_1 = x_2 + x_3 + \cdots + x_D, \tag{5.1}$$

This chapter is submitted as Vink, G., Pannekoek, J., & Van Buuren, S. Predictive ratio matching imputation of nested compositional data with semicontinuous variables.

where $x_2, \ldots, x_D$ take real positive values including 0. Suppose we have the following compositional data for $D = 3$ with missing values

x_1	x_2	x_3	x_4
32	10	15	7
18	0	18	0
22	6	3	–
14	0	–	–
–	–	16	–
8	22	–	4
30	–	–	–
–	5	12	15

For some of the missing values, it is possible to calculate the missing value. For example, row 3 yields $x_4 = 22 - (6 + 3) = 13$, and row 8 yields $x_1 = 5 + 12 + 15 = 32$.

If the data are inconsistent with the constraint (as in row 6) it follows that some values must be in error and imputations based on erroneous observations will also lead to erroneous or inconsistent data. Such errors need to be corrected before imputations can be generated. There is a vast literature on methods of detection and correction of errors in survey data - see e.g. (De Waal et al., 2011) -, but this topic is outside the scope of this paper. In this paper we assume that such inconsistencies have already been taken care of. The objective of the research is to generate multiple imputations for $x_2, \ldots, x_D$, assuming that x_1 is observed for every unit in the data. For situations where the total is missing (as in the fifth row) we require that this total has already been consistently filled in by some other method.

5.1.1 Existing approaches

Compositional data can be thought of as vectors of proportions, such that it holds that the D non-negative parts of any vector x sum up to a certain observed total. All the information about compositional data is encapsulated in the ratios between the components (Aitchison, 1986). Consequently, the proportions of the different parts of x obey

$$\frac{x_2}{x_1} + \frac{x_3}{x_1} + ... + \frac{x_D}{x_1} = 1 \tag{5.2}$$

which is equivalent to Equation (5.1), where

$$x_2 \geq 0, x_3 \geq 0, ..., x_D \geq 0. \tag{5.3}$$

Aitchison (1986) replaces the natural non-negativity condition of (5.3) by an assumption of strict positivity, thereby creating a formal definition of a D-dimensional simplex where all components must be larger than zero. In practice this may be undesirable or even unrealistic as bonafide zeros are observed. Examples of such bonafide zeros are, for instance, costs for investments or temporary personnel for businesses

that did not have such costs in a certain time frame, or rounded zeros in geology. Suggestions for dealing with zeros in compositional data sets have been made (Aitchison, 1986; Martín-Fernández et al., 2003; Aitchison et al., 2003; Palarea-Albaladejo et al., 2007).

Several strategies have been proposed to deal with missingness in compositional data sets. Martín-Fernández et al. (2003) advocate a multiplicative strategy for dealing with incomplete compositional data. In this multiplicative strategy, compositional parts are considered as proportions of a total. Missing proportions can be imputed and observed proportions have to be rescaled in concordance with the imputed proportions, thereby replacing the composition with a composition without missing values. As a result, observed values are altered to obtain a new composition that obeys all restrictions in problems where a fixed total is considered. Martín-Fernández et al. (2003) do not consider the 'best' value and do not introduce or suggest an imputation method.

Tempelman (2007) proposes stochastic and deterministic EM-based approaches, based on the Dirichlet distribution (also known as multivariate beta distribution), that can sometimes outperform non-parametric methods, such as nearest neighbor and random hot deck methods. Tempelman (2007) compares two Dirichlet imputation methods, one where expectations serve as imputations and one where random draws serve as imputions, to hot deck approaches that use random ratios or the ratio of the nearest neighbor, respectively.

Tempelman (2007) advices to use the expectation-based Dirichlet method when the aim of imputation is to preserve the population values, while hot deck approaches should be used when the distribution of the data is of interest. The EM-based approaches only model one equality restriction at once, which might be a limitation in practice, where variables can be subject to multiple equality restrictions.

Hron et al. (2010) propose a k-nearest neighbor (kNN) approach and an iterative model based imputation (IMI) technique that starts initially from the result of the proposed k-nearest neighbor procedure. The iterative model based imputation follows a sequential regression imputation strategy, where the regressions are carried out in an isometric log ratio (ilr) space (Egozcue et al., 2003). The reason for this ilr-transformation is that compositional data has no representation in Euclidean space, but rather in the simplex space. Back-transforming the ilr transformed data, however, results in updating both the original missings and the non-missing cells. Even though the ratios do not change, this may be undesirable when data analysis is considered, as observed values are usually assumed to be left intact. Hron et al. (2010) conclude that the proposed iterative model based imputation technique improves the initial kNN estimation and that it performs as least as good as existing imputation methods for compositional data. Hron et al. (2010) adapt the definitions from Aitchison (1986) and assume all components to be larger than zero. This assumption poses a restriction in practice, as compositional data may often be semicontinuous in many fields of statistics.

Another intuitive solution may be to sample the multivariate compositional distribution from a donor wherein all compositional parts are observed. There are a few

reasons that advocate the preference for a bivariate over a multivariate approach. First, sampling compositional distributions from a probable donor record works for missings in a single composition, but can not be generalized to nested compositions. For example, finding imputations by matching on an observed compositional distribution may lead to the sum of a nested compositional sum being smaller than the sum of its respective observed parts.

Second, sampling compositional distributions poses a much harsher restriction on the probable donor pool, especially when the number of compositional parts increases, as the number of observed records one can sample from decreases dramatically in the case of intermittent missingness. In general, when considering ratios, it can be said that a bivariate matching approach allows for the largest possible donor pool in the case of intermittent multivariate missingness. This research focuses on the problem of nested compositions and will therefor discuss bivariate imputation.

5.1.2 Properties

In our view an imputation method for compositional data must adhere to the following properties:

1. *Consistency:* Imputed and observed parts must sum to the total.
2. *Structure preserving:* The ratios between the components should be preserved and observed values must be left intact.
3. *Generality:* The number of missing components per record can vary and the component can be part of multiple nested compositions.
4. *Observed plausibility:* Imputed values must be within the range of plausible values.

The last property condition is particularly important when e.g. there are multiple semicontinuous variables in the composition, as imputations must either be attributed to the point mass or to the continuous part of the data, or when estimates at the edge of the distributions are considered. Not considering plausibility given the data at hand can lead to negative values being imputed in strictly positive data, for example.

5.2 Predictive Ratio Matching

We propose predictive ratio matching (PRM); a new, easily applicable bivariate hot deck imputation method for nested compositional data that makes use of the pairwise nature of the ratios between components.

5.2.1 Introduction of notation

Let X denote the matrix with n rows and p columns that contains the data. Further, x_{ij} denotes the i,jth element of X, where $i = 1,\ldots,n$ and $j = 1,\ldots,p$ and x_j denotes the jth column in X. Since we will consider imputation of the data in a pairwise manner, we also define $j' = 1,\ldots,j-1$ as the paired counterpart of j.

Together, j and j' can be used to form the lower triangular of the matrix of all possible combinations between the variables in X. Next, we define R to be a response indicator matrix of the same size as X, indicating 1 if x_{ij} is observed and 0 if x_{ij} is missing. We use $X_{jj'}$ to denote the columns j and j' in matrix X and we use $X_{-jj'}$ to denote matrix X except columns j and j'. Finally we denote observed elements with *obs* and missing elements with *mis*, such that e.g. X^{obs} and X^{mis} denote the observed and missing values in X, respectively.

5.2.2 A simple example

Let us consider X as a matrix containing the following 3-part compositional data

$$x_1 = x_2 + x_3 + x_4. \tag{5.4}$$

Let x_2 and x_3 be jointly missing for some cases. We know that the total amount $\sum X_{jj'} = x_j + x_{j'}$ to be distributed over x_2 and x_3 equals

$$\sum X_{32} = x_2 + x_3 = x_1 - x_4. \tag{5.5}$$

The distribution of $\sum X_{jj'}$ over x_j and $x_{j'}$ remains unknown. Assuming the ratio $\pi_{jj'} = x_{j'} / \sum X_{jj'}$ is the same in the observed and the missing data, we can solve this by finding imputed ratio $\pi^*_{jj'}$ through matching on the ratio $\pi_{jj'}$ from a probable donor record d, yielding

$$x^*_2 = \pi^*_{32} \sum X_{32}, \tag{5.6}$$

and its complement

$$x^*_3 = (1 - \pi^*_{32}) \sum X_{32}, \tag{5.7}$$

as imputations for x_2 and x_3, where π^*_{32} is the imputed ratio for pair 32 and comes from the distribution

$$\Pr(\pi^*_{jj'} | \pi_{jj'}, X_{-jj'}) \tag{5.8}$$

of donors with both x_j and $x_{j'}$ observed.

There are multiple strategies for obtaining an imputation $\pi^*_{jj'}$. Given the highly skewed nature of compositional data and given that imputed and observed values are allowed to be zero (as in semicontinuous data), we propose to use predictive mean matching (PMM) to impute the ratios. PMM is known to be very accurate in obtaining correct statistical inference and in retrieving the amount of zeros, even in highly skewed semicontinuous data (Vink et al., 2014).

5.2.3 Multivariate missingness in a single composition

If more than two components are missing in the composition, starting values can be computed that obey (5.4), whereafter the above approach can be iteratively used to obtain imputations for all pairs, each time selecting a probable donor ratio from the observed data and redistributing the amount over the current pair.

Suppose that x_2, x_3 and x_4 are jointly missing for some cases. We know that the total amount $\sum X_{jj'} = x_j + x_{j'}$ to be distributed over the unique pairs equals

$$\sum X_{32} = x_2 + x_3 = x_1 - x_4 \tag{5.9}$$

$$\sum X_{42} = x_2 + x_4 = x_1 - x_3 \tag{5.10}$$

$$\sum X_{43} = x_3 + x_4 = x_1 - x_2. \tag{5.11}$$

When x_2, x_3 and x_4 are jointly missing, these sums can not be deductively calculated from the observed data. This stresses the need for starting values. Any starting value will be sufficient as long as the compositional structure remains intact. For example, a simple strategy would be to divide a record's total amount of missingness $x_1 - \sum X_{-1}^{obs}$ equally over all the missing components in a record.

Once the starting values have been filled in, the approach introduced in 5.2.2 can be used to impute the missing unique pairs in the data in an iterative manner. For example, first pair X_{32} can be redistributed based on the ratio of a probable donor record. Because we found starting values that obey the composition, we can simply redistribute the amount over x_2 and x_3 based on the imputed ratio. We can continue this procedure for pair X_{42} and X_{43} and repeat it in the next iteration. We can continue iterating until convergence has been reached.

The MAR assumption implies that only the joint-missings for each pair have to be calculated and imputed. Partial missings in $X_{jj'}$ will be solved in another pair $X_{jj'}$ where x_j and $x_{j'}$ are both missing. Skipping over partially observed pairs is computationally convenient and ensures that the observed data remains intact.

It is wise to reorder the D-part composition based on the mean of the components. In that case ratios between adjacent variables are closer to 1. Variables with similar means are more efficiently imputed during the iterations, leading to better performance in datasets where some of the ratios between observed variable means are very large, or very small.

PRM algorithm

We require that starting values have been filled in and that any deductive imputation has been applied. Carry out the following steps for all $\binom{j}{2}$ unique pairs $X_{jj'}$.

1. Calculate $\pi_{jj'}^{obs}$ and replace all $\pi_{jj'}^{obs}$ that are not defined (when both x_j and $x_{j'}$ are 0) with $\pi_{jj'}^{obs} = 0.5$.
2. Impute $\pi_{jj'}^{*}$ by means of PMM with $\pi_{jj'}^{obs}$ conditional on $X_{-jj'}$.
3. For all joint missings in the pair $X_{jj'}$ distribute $\sum X_{jj'}$ following

$$x_{j'}^{*} = \pi_{jj'}^{*} \sum X_{jj'}$$
$$x_{j}^{*} = \sum X_{jj'} - x_{j'}^{*},$$

Repeat the above algorithm until convergence is reached. For multiple imputation do this $m \geq 2$ times, preferably in parallel with different random seeds, each time saving the completed dataset.

5.2.4 Nested compositions

Suppose that we have a nested composition, where x_4 is a combination of x_5 and x_6, such that

$$x_4 = x_5 + x_6, \tag{5.12}$$

resulting in the following extended data set from Section 5.1.

x_1	x_2	x_3	x_4	x_5	x_6
32	10	15	7	4	3
18	0	18	0	0	0
22	6	3	13	–	–
14	0	–	–	–	6
30	–	–	–	–	2
32	5	12	15	7	8

For the cases where x_4 is missing, the problem can be simplified to

$$x_1 = x_2 + x_3 + x_5 + x_6, \tag{5.13}$$

where x_4 is simply the sum of x_5 and x_6 and does not need to be imputed, but can be deductively calculated after x_5 and x_6 are imputed. This reduces the problem to a single composition, which can easily be solved by the proposed PRM algorithm from Section 5.2.3. The imputed value for x_4 can then be calculated as

$$x_4^* = x_5 + x_6, \tag{5.14}$$

yielding a solution where all imputed values obey all restrictions.

For the cases where x_4 is observed, the problem can be divided into the independent imputation problems

$$x_1 = x_2 + x_3 + x_4 \quad \text{and} \quad x_4 = x_5 + x_6. \tag{5.15}$$

Solving these separate compositions is also straightforward with the proposed PRM algorithm from Section 5.2.3. For example, an imputation for x_3 can be obtained by

$$x_3^* = (1 - \pi_{32}^*)(x_1 - x_4) \tag{5.16}$$

and imputations for x_5 are obtained by

$$x_5^* = \pi_{65}^*(x_4). \tag{5.17}$$

In both cases donors are drawn from within the compositional level of the missing values. If more than one pair in any of the (sub)compositions are missing, the above approach can be carried out iteratively.

Divide-and-conquer approach

In order to facilitate an approach that can handle compositions that are nested in other nested compositions, the compositional structure needs to be recorded. Let B denote a square binary matrix with p rows and p columns and zero diagonal, indicating 1 if a variable (column) is part of a compositional sum (rows). Column sums for B take value 0 for the highest level compositional sum only and take value 1 for all other variables. Row sums of B take values between 0 (when B_j is not a total) and $p-1$, indicating the number of parts in the decomposition of the j'th variable.

For ease of the argument let the rows and columns to be ordered with respect to the nesting levels from the highest level composition to the lowest level composition. For the previously used example in 5.2.4, the $p \times p$ matrix B takes the following form

$$B = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

We use b_c to denote the c'th row in B, with $c = p, \ldots, 1$. Note that $\sum b_c$ denotes the number of parts in the composition. Next, x_{ic} denotes the total of the c'th composition for row i in X, and r_{ic} indicates whether x_{ic} is observed or missing.

Extended algorithm

We require that starting values have been filled in and that any deductive imputation has been applied. We use R' and B' to denote "shadow copies" of R and B, respectively. Carry out the following steps for all rows c in B ($c = p, \ldots, 1$).

1. If $r_{ic} = 0$ and if $\sum b'_c \neq 0$
   a) Promote all nonzero elements and the diagonal in b'_c to row k of B', where k is the row for which $B_{kc} = 1$ and $k = 1, \ldots, p$.
2. If $r_{ic} = 1$ and if $\sum b'_c \neq 0$
   a) Impute the missing parts in the composition by means of PRM using missingness indicator R'.
   b) Set $r'_{ij} = 1$ if x_{ij} has been imputed in (a).
   c) Calculate unobserved nested totals (if any) in the current composition based on the imputed parts.
3. Repeat steps 1 and 2 for all rows c and afterwards set $R' = R$ and $B' = B$.
4. Reiterate the above steps until convergence is reached.

For multiple imputation execute the above algorithm $m \geq 2$ times, preferably in parallel with different random seeds, each time saving the completed dataset.

If the data is ordered based on the compositional hierarchy and if c moves from p to 1, nested compositions are considered before their respective higher level composition. The use of shadow copy B' makes it straightforward to consider nested compositions for which the total is not observed at a later stage in a higher level composition, where solving the nested compositional structure is possible. For example, shadow B' for the example from 5.2.4 takes the following form

$$B = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Simultaneously, making use of shadow R' prevents nested compositions for which the total has been observed and for which missing values have been imputed already, to be updated at a later stage. This ensures that the overall compositional structure remains intact and is computationally convenient, because only values that are 'still missing' have to be considered.

The proposed algorithm naturally handles any combination of nested compositions by determining for each composition which cases can be solved at the moment and which cases have to be promoted and solved within a higher level compositional problem. Ultimately, if none of the (multiply) nested compositional parts have observed totals, the composition is solved at the top-level composition by the PRM algorithm. For example, the PRM algorithm may yield the following multiple imputation ($m = 2$) solution for the case from Section 5.2.4.

$x1$	$x2$	$x3$	$x4$	$x5$	$x6$
32	10	15	7	4	3
18	0	18	0	0	0
22	6	3	13	6.1	6.9
14	0	2.7	11.3	5.3	6
30	7.1	11	11.9	9.9	2
32	5	12	15	7	8

$x1$	$x2$	$x3$	$x4$	$x5$	$x6$
32	10	15	7	4	3
18	0	18	0	0	0
22	6	3	13	6.1	6.9
14	0	2.7	11.3	5.3	6
30	0	28	2	0	2
32	5	12	15	7	8

Finding starting values for nested compositions

It is important to find acceptable starting values for the top-level composition that take the constraints put onto the values by the nested composition into account. Let g_i denote the amount that is missing in the top-level composition in x_i and let $z = 1, \ldots, D$ denote the D parts in the top-level composition. Finding suitable starting values for the top-level composition can be easily done by calculating, for each missing component x_{iz} in the top-level sum a starting value s, where

$$s_{iz} = \frac{\sum (x_{iz}^{mis})}{t_x} g_i, \tag{5.18}$$

with $\sum (x_{iz}^{mis})$ denoting the observed sum of the parts in the nested composition belonging to component x_{iz}^{mis} and t_x denoting the sum over all nested compositional parts for all missing components x_{ij}^{mis} in the top-level sum. If a component has no nested parts, or if all nested parts in a component are missing, $\sum (x_{ij}^{mis})$ should be set to 0.

For example, a solution for the missing parts in row 4 in the example from Section 5.2.4 that obeys the restriction would be

$$s_{43} = \frac{0}{6} \times 14 = 0 \quad \text{and} \quad s_{44} = \frac{6}{6} \times 14 = 14. \tag{5.19}$$

The above procedure ensures that starting values are at least as large as the sum of a components respective parts. It does so by taking a record's observed distribution and multiplying it with the unobserved mass, thereby leaving observed components in the top-level sum untouched. If starting values of the highest level components are properly calculated, starting values for nested compositions can be easily filled in. Please note that any starting value that obeys the compositional structure of the data will do for PRM and that the above procedure does not consider the best starting value.

5.3 Application

To illustrate PRM, we carried out a simulation study on a complete subset of a real dataset from official statistics, the Dutch Wholesaler Data (DWD) for 2007. The DWD dataset contains edited information on 1067 wholesalers for a set of cost statistics (a, e, g and h) that sum up to a set total x_1, leading to composition

$$x_1 = a + e + g + h, \tag{5.20}$$

where x_1 are the total operating costs and a, e, g and h represent the company depreciation, buying costs, personnel costs and other costs, respectively.

Component a is a single measure and component h forms a subcomposition with 21 parts, wherin the first component is summed over the following 3 parts, leading to nested composition

$$h_1 = h_2 + h_3 + h_4. \tag{5.21}$$

Finally, subcompositions g and e contain 9 and 5 parts, respectively. All parts over all subcompositions differ in mass and take proportions in the top-level sum x_1 ranging from .0001 to .9005.

All simulations are carried out in `R 3.0.2` (R Core Team, 2013) with 5 completed datasets, 2 iterations and 100 simulations. For the predictive mean matching step in the algorithm, `R`-package `mice` (version 2.18) Van Buuren and Groothuis-Oudshoorn (2011) is used.

5.3.1 Imposing missingness

We impose missingness in the DWD by means of two different missingness mechanisms: missing completely at random (MCAR) and missing at random (MAR). We distinguish between left-tailed MAR missingness and right-tailed MAR missingness. Particularly of interest is left-tailed MAR missingness, a realistic missingness mechanism in official business statistics, as larger companies tend to be always observed.

First, let $x_j = (x_j^{obs}, x_j^{mis})$ denote the variable (compositional part) to be made incomplete, with *obs* and *mis* indicating the observed and missing sections of x_j, respectively. Second, let R denote the missingness indicator with $R = 0$ when x_j is missing and $R = 1$ when x_j is observed. We create MAR missingness in each compositional part x_j, based on x_1 in our samples according to the following mechanism

$$P(R = 0|x_j^{obs}, x_j^{mis}, x_1) = P(R = 0|x_j^{obs}, x_1),$$

by using a random draw from a binomial distribution of the same length as x_j and of size 1 with missingness probability equal to the inverse logit

$$P(R = 0) = \frac{\exp(\alpha)}{(1 + \exp(\alpha))}.$$

In the case of left-tailed MAR missingness, $\alpha = (-\bar{x}_1 + x_1)/\sigma_{x_1}$ gives 50 percent missingness, where σ_{x_1} indicates the standard deviation of variable x_1. For right-tailed MAR missingness, this can be achieved by choosing $\alpha = (\bar{x}_1 - x_1)/\sigma_{x_1}$. Adding or substracting a constant moves the sigmoid curve in the horizontal direction, which results in different missingness proportions. Simulations are carried out with 15 and 25 percent univariate missingness. The number of completely observed rows varies between 0 and 15.

Bivariate missingness

PRM is a bivariate imputation algorithm that uses the relation between observed values in each pair. As a consequence, the actual missing data problem is larger than the amount of univariate missingness would suggest. For example, under MCAR, 15 percent univariate missingness leads to $15 + (.15 * 75) = 27.75$ percent bivariate missingness and 25 percent univariate missingness results in $25 + (.25 * 75) = 43.75$ percent bivariate missingness.

5.3.2 Evaluation

Performance of the method is evaluated by looking at studying unbiasedness of the estimate of the population mean (weighted sums), coverage rates of the 95% confidence interval of the estimates, ratio's between components after imputation and the amount of zeros that is recovered by the algorithm.

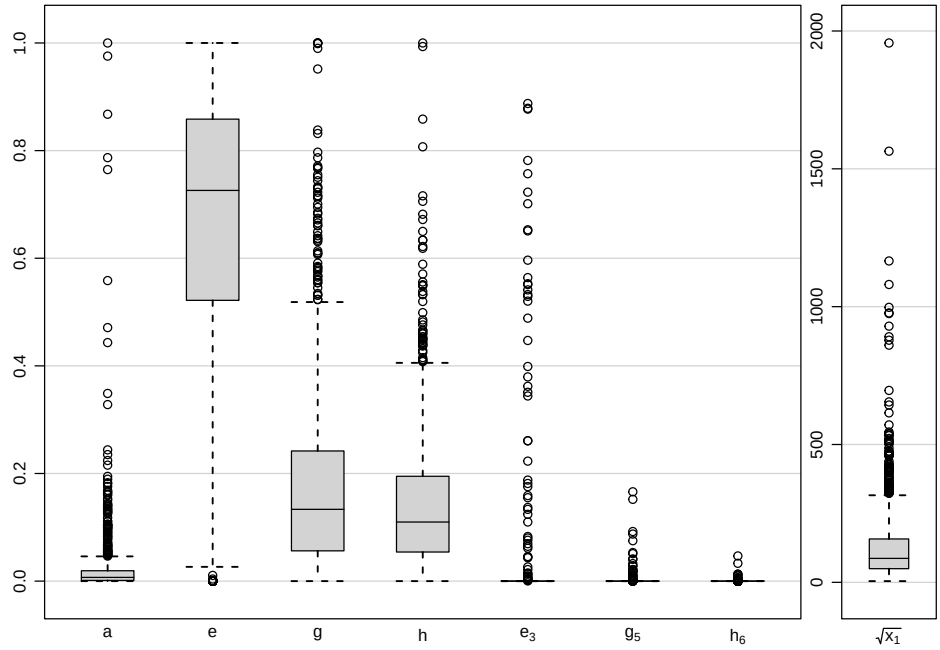

Fig. 5.1. Population boxplots for nested sums a, e, g and h and their respective smallest nested parts e_3, g_5 and h_6. Displayed are the distribution of proportions the respective variables take in the top-level sum x_1 and, next to it, the distribution of $\sqrt{x_1}$.

5.4 Results

The extreme skewness in the distribution of the variables are displayed in Figure 5.1. The skewed nature of the variables poses a difficulty for any imputation algorithm. The hot deck method used in the PRM algorithm is highly non-parametric and does not directly depend on any distributional assumption, resulting in imputations that are very close to the distribution of the observed data (Vink et al., 2014).

5.4.1 Means

Table 5.1 shows simulation results for estimates of the population mean. It becomes clear that PRM under left-tailed MAR missingness outperforms the results of PRM under MCAR and right-tailed MAR missingness, thereby showing very low bias across all simulation conditions. Simulation results for left-tailed MAR missingness are very stable across the simulation conditions. Performance under MCAR decreases slightly when more missingness is introduced, but overall performance under MCAR can still be considered to be good. In all simulation conditions, the performance of imputation

Table 5.1. Means of the variables in DWD. Displayed are the proportions the mean of the variables take in sum x_1, for the true population values and for imputations under 3 different missingness mechanisms for 15 percent and 25 percent intermittent univariate missingness.

		15% missingness			25 % missingness		
	true	mcar	left	right	mcar	left	right
x_1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
a	0.005	0.006	0.005	0.015	0.007	0.005	0.016
h	0.055	0.057	0.056	0.258	0.066	0.057	0.276
g	0.039	0.041	0.040	0.159	0.047	0.041	0.157
e	0.901	0.896	0.900	0.568	0.881	0.897	0.551
h_1	0.001	0.001	0.001	0.050	0.002	0.001	0.043
h_2	0.001	0.001	0.001	0.017	0.001	0.001	0.015
h_3	0.000*	0.000*	0.000*	0.014	0.001	0.000*	0.014
h_4	0.000*	0.000*	0.000*	0.019	0.000*	0.000*	0.014
h_5	0.004	0.004	0.004	0.016	0.005	0.004	0.018
h_6	0.000*	0.000*	0.000*	0.006	0.000*	0.000*	0.005
h_7	0.000*	0.000*	0.000*	0.007	0.000*	0.000*	0.005
h_8	0.001	0.001	0.001	0.011	0.001	0.001	0.009
h_9	0.002	0.002	0.002	0.012	0.002	0.002	0.015
h_{10}	0.000*	0.000*	0.000*	0.006	0.000*	0.000*	0.004
h_{11}	0.003	0.004	0.003	0.015	0.004	0.004	0.014
h_{12}	0.001	0.002	0.001	0.011	0.002	0.001	0.017
h_{13}	0.012	0.012	0.012	0.018	0.012	0.012	0.019
h_{14}	0.001	0.001	0.001	0.010	0.002	0.001	0.009
h_{15}	0.001	0.001	0.001	0.009	0.001	0.001	0.009
h_{16}	0.006	0.006	0.006	0.017	0.008	0.006	0.026
h_{17}	0.001	0.001	0.001	0.009	0.001	0.001	0.010
h_{18}	0.003	0.003	0.003	0.013	0.004	0.004	0.013
h_{19}	0.004	0.004	0.004	0.010	0.004	0.004	0.012
h_{20}	0.006	0.006	0.006	0.015	0.006	0.005	0.022
h_{21}	0.003	0.003	0.003	0.010	0.003	0.003	0.011
h_{22}	0.006	0.006	0.006	0.013	0.007	0.006	0.015
g_1	0.028	0.028	0.028	0.039	0.029	0.028	0.043
g_2	0.003	0.003	0.003	0.013	0.004	0.003	0.013
g_3	0.002	0.002	0.002	0.013	0.003	0.002	0.011
g_4	0.000*	0.001	0.000*	0.014	0.001	0.001	0.010
g_5	0.000*	0.000*	0.000*	0.010	0.001	0.000*	0.006
g_6	0.003	0.003	0.003	0.020	0.004	0.003	0.025
g_7	0.001	0.002	0.001	0.028	0.002	0.001	0.030
g_8	0.000*	0.000*	0.000*	0.007	0.001	0.000*	0.005
g_9	0.002	0.002	0.002	0.016	0.002	0.002	0.013
e_1	0.866	0.854	0.863	0.511	0.824	0.858	0.477
e_2	0.005	0.007	0.005	0.013	0.011	0.006	0.016
e_3	0.004	0.006	0.005	0.010	0.009	0.005	0.014
e_4	0.015	0.016	0.016	0.014	0.019	0.016	0.020
e_5	0.011	0.012	0.011	0.020	0.018	0.012	0.024

* indicates values smaller than .0005

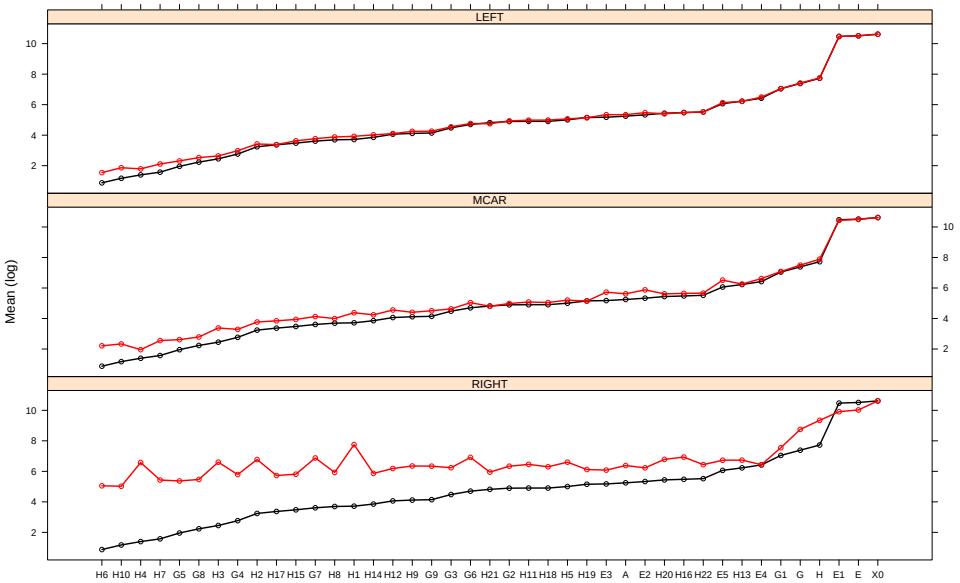

Fig. 5.2. Simulation results for 25 percent univariate missingness. Please note that means are log-transformed. Red lines represent completed data, black lines represent observed data.

under right-tailed MAR missingness is worse than for MCAR and left-tailed MAR missingness. As would be expected, variables with smaller mean values show larger relative bias than variables with larger mean values. Thus, in absolute sense, the bias is always small. However we need care under left-tailed MAR and MCAR when we study the relative contributions of small parts. Figure 5.2 clearly display this finding. The slight underperformance of MCAR and more pronounced loss in performance for right-tailed MAR, compared to left-tailed missingness, is caused by the distribution of the pairwise relations in the data and the missingness creation itself. For example, under MCAR each cell has the same probability to be missing, resulting in larger values and smaller values being 'removed' with the same likelihood. However, the algorithm iteratively redistributes the sum of two pairs of variables over the pairs, based on their individual mass. The redistribution of amounts over cases with very large (or very small) ratios leads to greater relative impact on the mean of the smaller value, thus causing bias in this particular setting.

The log-transformation of the means in Figure 5.2 makes this property more apparent. With right-tailed MAR missingness, larger values are more likely to be missing, resulting in a more equal redistribution of missing amounts over the missing cells in each case. In contrast, a left-tailed MAR missingness mechanism creates more missing data in the smaller values, but leaves the larger values intact. As a result, there is a larger relative bias in the smaller values. However, the absolute bias for all variables

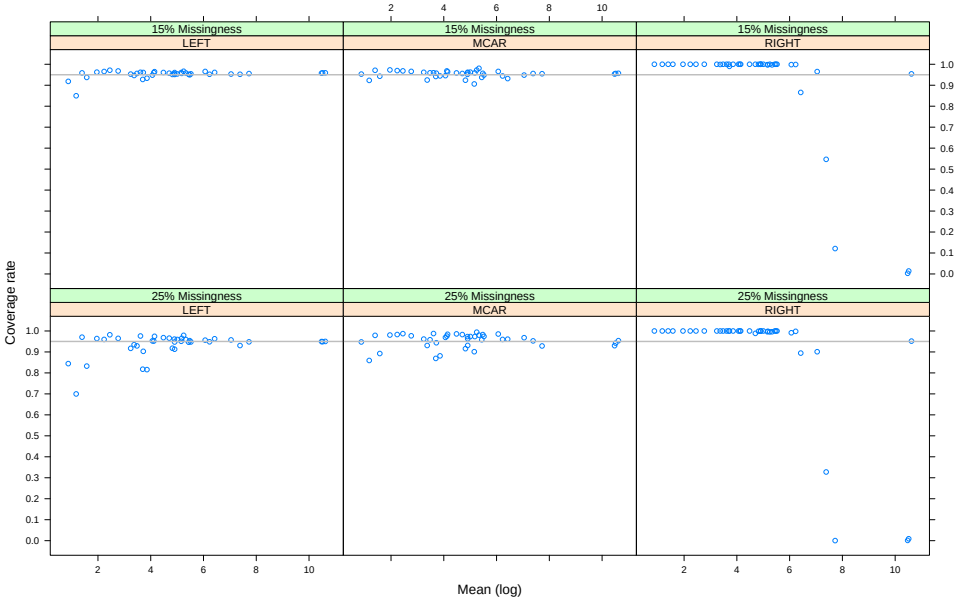

Fig. 5.3. Coverage rates plotted against the mean for different missingness mechanisms with different amounts of missingness. Please note that means are log-transformed

is very low for any missingness amount. Especially for economical and business statistics, the bias towards larger values for smaller posts is a surprisingly useful property, as the missingness in these fields is almost exclusively considered to be left-tailed.

5.4.2 Coverage and confidence interval width

There is a relation between coverage rates and the relative importance (or impact) the variables have in the composition. This relation is shown in Figure 5.3. For 15 and 25 percent univariate missingness, coverage rates are very close to the nominal level.

The performance of PRM for left-tailed MAR missingness and MCAR is similar. However, when right-tailed MAR missingness is considered, the algorithm tends to undercover the confidence interval for components with large means. In general, for smaller amounts of missingness under MCAR or left-tailed MAR, variables with less mass are more prone to undercoverage than variables with more mass. This is the result of the redistribution of missing amounts over the incomplete pairs (see Section 5.4.1 for a more detailed explanation).

The relation between mean and coverage rate becomes less apparent in the case of 50 percent univariate missingness (not shown), due to grand loss of information. Interestingly, for large amounts of missingness, left-tailed MAR missingness remains

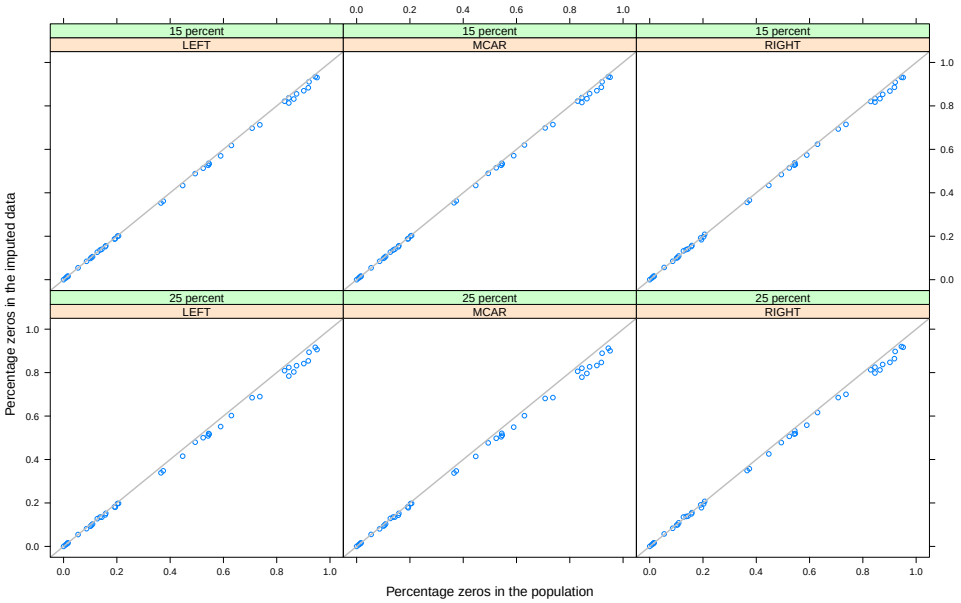

Fig. 5.4. The proportion of zeros after imputation, plotted against the true proportion of zeros in the population. Shown are estimates under different missingness mechanisms with different amounts of missingness.

able to sufficiently cover the confidence interval of the very large compositional parts, as opposed to the other missingness mechanisms.

5.4.3 Zeros

Predictive mean matching is known to be very accurate in retrieving the amount of zeros in semicontinuous data (Vink et al., 2014). Since predictive mean matching is used in PRM as a donor selection approach, it is expected that the percentage of zeros after imputation will not be very different from the population values. Figure 5.4 shows the bias of the amount of zeros after imputation plotted against the true population values.

From Figure 5.4 it follows that the performance of PRM in retrieving the amount of zeros after imputation depends mostly on the true amount of zeros in the population and in to a lesser extend on the severity of the missingness. However, with increasing amounts of zeros, accuracy naturally decreases. In situations with 15 percent univariate missingness bias is negligible. For 25 percent univariate missingness, bias is acceptable, especially as larger biases only occur in variables with at least eighty percent zeros; a situation that poses harsh restrictions on the amount of usable information in the data.

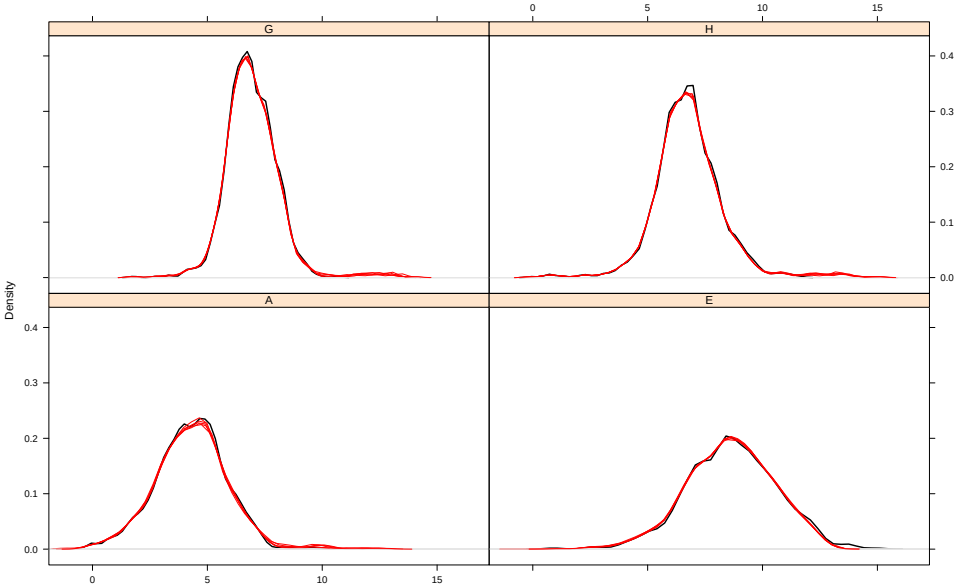

Fig. 5.5. Densities of the parts of the highest level composition (x_1) for 25 percent right-tailed MAR missingness. The black line shows the population density, while red lines show the densities of the 5 multiply imputed variables. Variables are highly skewed and have been log transformed for plotting.

There is no clear difference in performance of PRM over different missingness mechanism. This can be explained by the data-driven hot deck nature of the approach.

5.4.4 Distributional shapes

Figure 5.5 displays the densities of the 4 highest level compositional parts for 25 percent right-tailed MAR missingness; a missingness mechanism that poses a particular constraint on the performance of PRM. We can see that imputed densities are very close to the population densities, although some larger values are imputed for some variables. This is due to the nature of the missingness and the bivariate approach that is used in PRM.

5.4.5 Convergence of the algorithm

Figure 5.6 displays the means and the standard deviation for the multiple imputation chains. It seems that the algorithm converges within a couple of iterations. When viewed over 100 iterations, the plot does not show a clear trend at the final iteration. Occasionally, one of the multiple imputation chains yields higher mean and standard

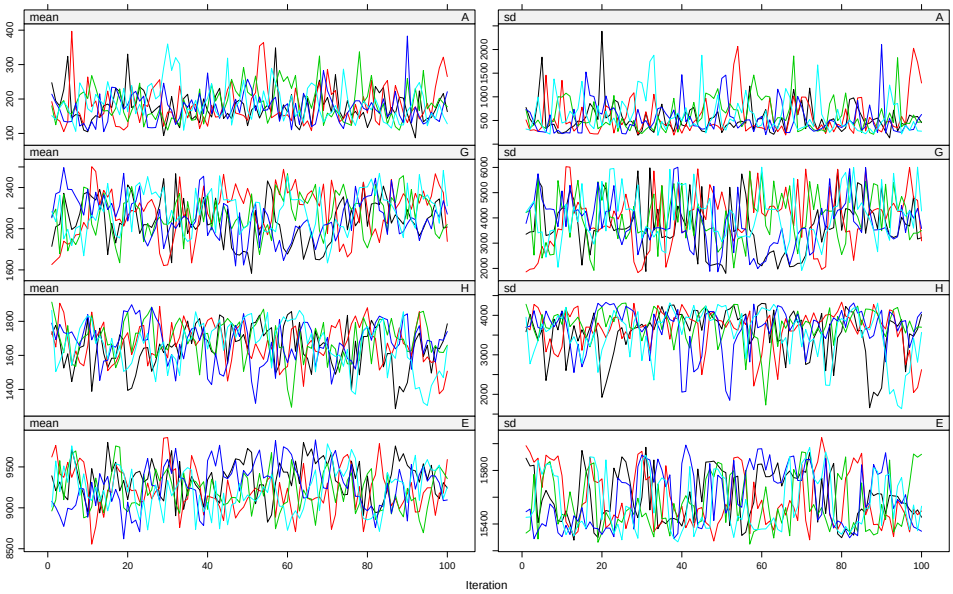

Fig. 5.6. Plots for the mean and standard deviation for nested sums a, g, h and e over 100 iterations. The different colors represent the $m = 5$ multiple imputation chains

deviation than the other chains, indicating that at least one larger value is imputed in the respective variable at that particular iteration. This occurs mainly in the variables that contribute less to the compositional sum.

5.5 Conclusion

We introduced and evaluated PRM, a bivariate imputation approach for nested compositional data. PRM emerges as a very effective imputation approach for nested compositional data and can handle the skewed semicontinuous variables in the simulation dataset.

PRM requires the top level sum to be observed. In case of an incomplete top level sum, standard imputation approaches could be used to first impute plausible values for the incomplete top level sum. Naturally, such imputations need to obey any edit restrictions associated with that particular sum.

As an alternative to the proposed donor selection, the ratio $x_j/x_{j'}$ could be used, but this poses practical problems with regard to symmetry. For example, ratio $x_j/x_{j'}$ has a different relation with $X_{-jj'}$ than ratio $x_{j'}/x_j$, whereas the proposed $\pi_{jj'}$ and $1 - \pi_{jj'}$ have an equal relation with $X_{-jj'}$, albeit of opposite sign.

Imputations generated by PRM will take edit restrictions associated with each separate composition into account while ensuring that imputations simultaneously obey

the overall nested compositional structure of the data. Imputing the ratios addresses the problem of incomplete compositional data out of the simplex space. This is very convenient and allows for (large amounts of) zeros in the data, without the need for any transformations or a post-hoc fix. Further, because of the hot deck nature of the method, no specific distributional assumptions have to be made. Finally, the proposed PRM method leaves the observed data values and ratios between components intact.

A comparison between imputation approaches for single compositions remains topic for future research, but is beyond the scope of this research.

Part III

Pooling imputations

6

Pooling multiple imputations when the sample happens to be the population

Summary. Current pooling rules for multiply imputed data assume infinite populations. In some situations this assumption is not feasible as every unit in the population has been observed, potentially leading to over-covered population estimates. We simplify the existing pooling rules for situations where the sampling variance is not of interest. We compare these rules to the conventional pooling rules and demonstrate their use in a situation where there is no sampling variance. Using the standard pooling rules in situations where sampling variance should not be considered, leads to overestimation of the variance of the estimates of interest, especially when the amount of missingness is not very large. As a result, populations estimates are over-covered, which may lead to a loss of statistical power. We conclude that the theory of multiple imputation can be extended to the situation where the sample happens to be the population. The simplified pooling rules can be easily implemented to obtain valid inference in cases where we have observed essentially all units and in simulation studies addressing the missingness mechanism only.

6.1 Background

Missing data are an ubiquitous problem in medical research. The occurrence of missing data often has an influence on the precision of estimates and may even lead to biased estimates and incorrect statistical inferences. A straightforward approach to obtain valid inference on incomplete data is multiple imputation. With multiple imputation the missing data problem is solved before the analysis takes place and each missing value is imputed $m \geq 2$ times, leading to m completed datasets. These m completed datasets are then analyzed separately and their complete data estimates are combined using Rubin's rules (Rubin, 1987).

This chapter is published as Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv preprint arXiv:1409.8542.*

Current methodology on pooling estimates based on multiply imputed data assumes the data are sampled from infinite populations. In some cases we have data on all units, e.g. rare conditions in medical research and registers in official statistics, and sampling variation plays no role. Yet, even though all units are observed, there may be missing data that affect the precision of the estimates of interest. In such situations, assuming an infinite population may overestimate the variance of the estimates. As a result, confidence intervals are longer than needed, leading to a loss of statistical efficiency.

This note suggests the use of simplified pooling rules that only account for the variation caused by the mechanism that created the missing data.

Methods

Rubin (1987, p. 76) defined Q as the quantity of interest (possibly a vector) and U as its variance. With multiple imputation, m complete data estimates can be averaged as

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l \tag{6.1}$$

where $\hat{Q}_l$ is an estimate of Q from the l-th imputed data set. Let $\bar{U}_l$ be the estimated variance-covariance matrix of $\hat{Q}_l$. The complete data variances of Q can be combined by

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m}\bar{U}_l. \tag{6.2}$$

The variance between the complete data estimates can be calculated as

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})'(\hat{Q}_l - \bar{Q}). \tag{6.3}$$

The total variance of $(Q - \bar{Q})$ is defined as

$$T = \bar{U} + B + B/m. \tag{6.4}$$

For populations for which all units are recorded, the average complete data variance $\bar{U}$ of Q equals zero - there is no sampling variation - and the total variance of $(Q - \bar{Q})$ simplifies to

$$T = B + B/m. \tag{6.5}$$

As a consequence, the relative increase in variance due to nonresponse equals

$$r = (1 + m^{-1})B/\bar{U} = \infty, \tag{6.6}$$

and the degrees of freedom ν can be set to

$$\nu = (m-1)(1 + r^{-1})^2 = m - 1. \tag{6.7}$$

Simulation

We created a finite population with $N = 1000$ members by drawing 1000 independent realizations from the multivariate normal distribution with means

$$\mu = \begin{matrix} X \\ Y_1 \\ Y_2 \end{matrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \tag{6.8}$$

and covariance structure

$$\Sigma = \begin{matrix} \\ X \\ Y_1 \\ Y_2 \end{matrix} \begin{matrix} \begin{matrix} X & Y_1 & Y_2 \end{matrix} \\ \begin{pmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix} \end{matrix}, \tag{6.9}$$

where X is a completely observed covariate and Y_1 and Y_2 are made incomplete by randomly deleting values with probabilities that vary between 0.1 and 0.95.

Data imputations are performed with `mice` (Van Buuren and Groothuis-Oudshoorn, 2011, version 2.21) in `R` (R Core Team, 2013, version 3.0.2) with Bayesian linear regression imputation (`mice.impute.norm`) as the imputation method and 10 iterations for the algorithm to converge. The quantities of scientific interest were the means of Y_1 and Y_2. The true values were calculated as the sample means before deletion.

Results

The results over 10000 simulations are shown in Table 6.1. It is clear that excluding the sampling variation in $\bar{U}$ from the confidence interval calculation leads to proper coverage of the 95% confidence interval of the mean. Taking $\bar{U}$ into account when considering completely observed populations, leads to overcoverage. Not surprisingly, this overcoverage becomes less apparent when the fraction of information missing due to nonresponse approaches 1.

With increasing missingness, the role of the between variance B in the total variance T in Rubin's rules becomes increasingly more important and the relative contribution of $\bar{U}$ in T decreases. Due to the increase in r, the resulting degrees of freedom ν approach $m - 1$. Eventually, when all data are missing, sampling variation $\bar{U}$ disappears and both pooling approaches become equivalent (see Figure 6.1).

All estimates are unbiased. As expected, the conventional pooling rules overestimate the total variance in the datasets covering the entire population, leading to overcoverage (c.f. Table 6.1). In contrast, the simplified pooling rules yield consistently a coverage of 95% of the 95 percent confidence interval.

Conclusions

This note illustrates that sharper inferences are possible in situations where the entire population has been observed, and all variation stems for the missing data. Our simulations show that the simplified pooling rules yield variance-covariance estimates that lead to shorter confidence intervals with correct statistical properties.

Although the adaptation to the conventional pooling rules is small (and may be considered mathematically trivial), we are not aware of any work actually applying simplified pooling. There are several instances where the simplified rules may be of practical interest. First, in situations where essentially all units are observed but missingness has its influence on the precision of estimates, multiple imputation can be utilized to obtain sharper inferences when the proposed pooling rules are used to obtain inference. Such applications can be found throughout many scientific fields, such as medicinal sciences, official statistics and big data applications.

Another useful application can be found in simulation studies involving the evaluation of imputation approaches. For the last decades, sampling variation has been an essential part of the evaluation of multiple imputation approaches. When infi-

Table 6.1. Coverage of the mean. Average results over 10,000 simulations for two variables with varying percentage of missingness. Results are shown for pooling rules for finite populations (simplified rules) and for the pooling rules as defined by Rubin (conventional rules).

		conventional rules					simplified rules			
	%mis	r	ν	fmi	ciw	cov	r	ν	ciw	cov
Y_1	10	0.13	330.36*	0.12	0.13	1.000	∞	4	0.06	0.949
	20	0.30	775.54	0.23	0.14	1.000	∞	4	0.09	0.950
	30	0.51	401.36	0.33	0.15	0.999	∞	4	0.11	0.950
	40	0.80	400.58	0.44	0.17	0.994	∞	4	0.14	0.951
	50	1.20	57.82	0.53	0.19	0.988	∞	4	0.18	0.951
	60	1.81	31.55	0.62	0.23	0.976	∞	4	0.22	0.950
	70	2.82	20.48	0.71	0.27	0.969	∞	4	0.27	0.953
	80	4.78	14.18	0.80	0.35	0.958	∞	4	0.35	0.952
	90	10.84	6.30	0.89	0.53	0.950	∞	4	0.54	0.951
	95	22.79	5.44	0.94	0.79	0.948	∞	4	0.80	0.951
Y_2	10	0.13	6483.96	0.12	0.14	1.000	∞	4	0.06	0.948
	20	0.29	1028.42	0.23	0.15	1.000	∞	4	0.09	0.950
	30	0.50	261.06	0.33	0.16	0.999	∞	4	0.12	0.948
	40	0.78	167.69	0.43	0.18	0.995	∞	4	0.15	0.949
	50	1.17	90.94	0.52	0.20	0.988	∞	4	0.18	0.951
	60	1.78	68.54	0.62	0.24	0.978	∞	4	0.23	0.949
	70	2.72	19.66	0.70	0.29	0.966	∞	4	0.28	0.949
	80	4.71	11.71	0.80	0.37	0.960	∞	4	0.37	0.951
	90	10.61	8.24	0.89	0.56	0.949	∞	4	0.57	0.949
	95	22.40	5.23	0.94	0.83	0.944	∞	4	0.84	0.947

* Calculated cf. Barnard and Rubin (1999) because occasionally $r \approx 0$.

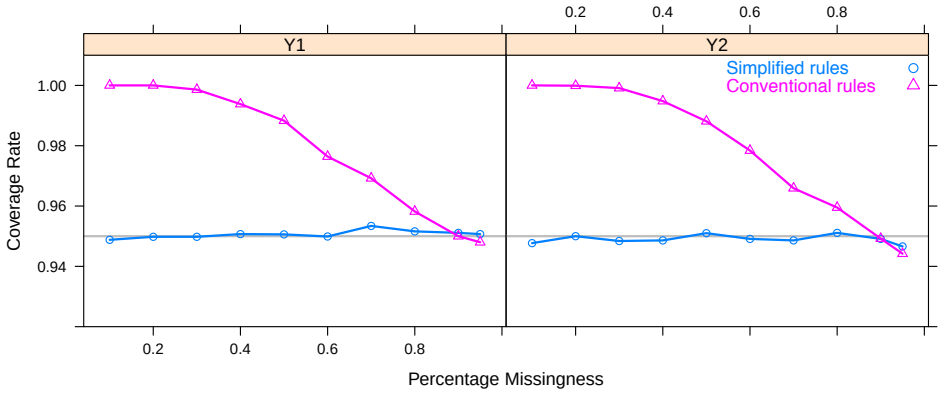

Fig. 6.1. Simplified and conventional pooling rules compared. Displayed are coverage rates for different missingness rates when assuming finite (simplified rules) or infinite (conventional rules) populations.

nite populations cannot be assumed during such evaluations, design-based simulation strategies are often used to properly account for sampling variation. However, in order to obtain information about a method's ability to handle the missing data problem, or to objectively compare methods on their ability to correct for missingness, it is not necessary to take sampling variation into account. After all, we are interested only in the missing data mechanism, and are not considering the noise induced by the sampling mechanism for evaluation in such studies. Moreover, not having to consider the sampling mechanism makes the generation of simulation data much more straightforward, especially when generating intricate multivariate data structures.

References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):273–291.

Acara (2014). *Guide to understanding 2013 index of community socio-educational advantage (ICSEA) values*. Australian curriculum assesment and reporting authority, Sydney, Australia. Last accessed on Sep 30, 2014. Available from: `http://www.acara.edu.au/verve/_resources/Guide_to_understanding_2013_ICSEA_values.pdf`.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall.

Aitchison, J., Kay, J., et al. (2003). Possible solution of some essential zero problems in compositional data analysis. In *Conference paper for the CoDaWORK workshop on Compositional data analysis, Girona, September 2003*.

Alfons, A., Templ, M., and Filzmoser, P. (2010a). Applications of statistical simulation in the case of eu-silc: Using the r package simframe. *Journal of Statistical Software*, 37(3):17.

Alfons, A., Templ, M., and Filzmoser, P. (2010b). An object-oriented framework for statistical simulation: The r package simframe. *Journal of Statistical Software*, 37(3):1–36.

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics*, 24(1-2):3–61.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53(1):57–74.

Barnard, J. and Rubin, D. B. (1999). Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and for the Alzheimer's Disease Neuroimaging Initiative* (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, Advance online publication. Available from: `http://smm.sagepub.com/content/early/2014/03/31/0962280214521348.abstract`.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675.

Brand, J. P. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus MC: University Medical Center Rotterdam.

Chambers, R. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Chen, H. Y., Xie, H., and Qian, Y. (2011). Multiple imputation for missing values through conditional semiparametric odds ratio models. *Biometrics*, 67(3):799–809.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Duan, N., Manning, W., Morris, C., and Newhouse, J. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, pages 115–126.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., and Knudtson, M. L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, 55(2):184–191.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197.

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, 42(4):679–694.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *NBER Chapters*, pages 120–137.

Heeringa, S., Little, R., and Raghunathan, T. (2002). *Survey Nonresponse*, chapter Multivariate Imputation of Coarsened Survey Data on Household Wealth, pages 357–371. Wiley.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Psychology Press, 2 edition.

Hron, K., Templ, M., and Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54(12):3095–3107.

Javaras, K. and Van Dyk, D. (2003). Multiple imputation for incomplete data with semicontinuous variables. *Journal of the American Statistical Association*, 98(463):703–715.

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. Technical report, mimeo, Board of Governors of the Federal Reserve System.

Koller-Meinfelder, F. (2009). *Analysis of Incomplete Survey Data–Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching.* PhD thesis, Otto-Friedrich-Universität Bamberg.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624–632.

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

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data.* Wiley-Interscience.

Manning, W., Morris, C., Newhouse, J., Orr, L., Duan, N., Keeler, E., Leibowitz, A., Marquis, K., Marquis, M., and Phelps, C. (1981). A two-part model of the demand for medical care.: preliminary results from the health insurance study. *Economics and Health Economics. Amsterdam: North-Holland.*

Martín-Fernández, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1):75.

Olkin, I. and Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2):448–465.

Olsen, M. and Schafer, J. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Oudshoorn, C. G. M., Buuren, S., and Rijckevorsel, J. L. A. (1999). *Flexible multiple imputation by chained equations of the AVO-95 survey.* TNO Prevention and Health Leiden.

Palarea-Albaladejo, J., Martín-Fernández, J., and Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7):625–645.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from: `http://www.R-project.org/`.

Raghunathan, T., Solenberger, P., and Van Hoewyk, J. (2002). IVEware: imputation and variance estimation software. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.*

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Raghunathan, T. E. and Siscovick, D. S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45(3):335–352.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4:227–241.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *Stata Journal*, 5:527–536.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

Sargasso.nl. De haagse twitter stolp [online]. (2012) [cited March 2012]. Available from: `https://docs.google.com/spreadsheet/ccc?key=0AhASywKnYQZqdEhUc2Z6Mk4tTm1TbXFnVTZzZ1d3aVE#gid=0`.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Schafer, J. and Olsen, M. (1999). Modeling and imputation of semicontinuous survey variables. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference*.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4):425–446.

Seaman, S., Bartlett, J., and White, I. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, 12(1):46.

Siddique, J. and Belin, T. (2007). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, 27(1):83–102.

Su, Y., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45(2).

Tempelman, C. (2007). *Imputation of restricted data*. PhD thesis, Doctorate thesis, University of Groningen.

Templ, M., Kowarik, A., and Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55:2793–2806.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC.

Van Buuren, S., Boshuizen, H., and Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67.

Van Buuren, S. and Oudshoorn, C. (2000). Multivariate imputation by chained equations: Mice v1. 0 user's manual. *The Netherlands: TNO Report PG/VGZ/00.038. Netherlands Organization for applied scientific research.*

Van Buuren, S. and van Rijckevorsel, J. L. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, 57(4):567–580.

Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology*, 60(2):315–338.

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

Vink, G. and Van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. arXiv:1409.8542 [math.ST].

Von Hippel, P. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291.

White, I., Royston, P., and Wood, A. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Yu, L., Burton, A., and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semicontinuous data. *Statistical Methods in Medical Research*, 16(243).

Zhao, J. H. and Schafer, J. L. (2013). *pan: Multiple imputation for multivariate panel or clustered data.* R package version 0.9.

List of Figures

List of Tables

Samenvatting

Ontbrekende data vormen een alomtegenwoordig probleem waar de meeste wetenschappers of onderzoekers vroeg of laat mee te maken krijgen. Een goed voorbeeld kan men vinden in onderzoek waar men gebruik maakt van vragenlijsten. Niet zelden slaan deelnemers aan dergelijke onderzoeken één of meerdere vragen over, met als resultaat dat de data niet compleet geobserveerd zijn. Dit vormt een probleem omdat de meeste statistische analyses veronderstellen dat de data compleet zijn.

Een veelgebruikte ad hoc oplossing voor het analyseren van incomplete data ligt in het negeren van de ontbrekende data. Echter, ontbrekende waarden kunnen niet zonder meer worden genegeerd. Immers, berekeningen op de geobserveerde data alleen kunnen een vertekend beeld geven wanneer er een reden is voor het ontbreken van (delen van) de data. Zelfs als er geen reden zou zijn voor het ontbreken van bepaalde waarden, resulteert een analyse van de geobserveerde data in een lager aantal respondenten dat gebruikt kan worden. Als gevolg hiervan is er een lagere kans op het vinden van een onderzoekseffect wanneer dit effect aanwezig is (statistische power) en zijn standaardfouten vertekend. Met andere woorden, de significantieniveaus (p-waarden) die gemeengoed zijn in wetenschappelijk onderzoek en de achterliggende conclusies zijn in essentie fout wanneer men de ontbrekende data simpelweg negeert.

Om incomplete data op een juiste manier te kunnen analyseren, zijn er twee algemeen geaccepteerde mogelijkheden. De eerste oplossing ligt in analyses die gebruik maken van schattingsmethoden die met ontbrekende data kunnen omgaan. Hierbij kan worden gedacht aan technieken zoals maximum likelihood, wegen en volledige Bayesiaanse schattingstechnieken. Het incomplete data probleem wordt hierbij 'opgelost' tijdens de analyse. Men kan echter ook het analyseproces en het probleem van ontbrekende data loskoppelen. Dit wordt gedaan in wat men imputatie noemt. Ontbrekende waarden worden geïmputeerd (ingevuld) en de gecompleteerde dataset kan vervolgens worden geanalyseerd met behulp van standaard analysetechnieken.

Wanneer men de ontbrekende data slechts eenmaal zou imputeren, wordt de onzekerheid rond de invullingen niet in de geïmputeerde data weerspiegeld en dienen er specifieke methoden voor het schatten van de standaardfouten te worden gebruikt. Een meer flexibele techniek om de onzekerheid omtrent de ingevulde waarden mee te

nemen is multiple imputation (MI). Met MI wordt iedere ontbrekende waarde $m \geq 2$ maal ingevuld om zo m gecompleteerde datasets te verkrijgen. Ten minste twee imputaties zijn nodig om de onzekerheid omtrent de invullingen te beschouwen, maar doorgaans is het aan te raden om een groter aantal imputaties te kiezen. De m datasets kunnen vervolgens worden geanalyseerd als ware het compleet geobserveerde datasets en de m analyseresultaten kunnen worden samengevat in één enkele gevolgtrekking.

In dit proefschrift worden enkel op MI gebaseerde methoden behandeld. De keuze voor MI is gebaseerd op de volgende argumenten. Ten eerste, MI wordt in toenemende mate meer populair en is zonder twijfel één van de meest gebruikte methoden voor het omgaan met nonrespons. Het aantal boeken en conferenties waarin MI wordt beschouwd neemt dan ook snel toe. Een mogelijke verklaring voor de populariteit van MI is de relatieve eenvoud waarmee conclusies op basis van MI kunnen worden verkregen en kunnen worden uitgelegd. Dit zijn aantrekkelijke eigenschappen die in het bijzonder toegepaste onderzoekers zullen aanspreken.

Een tweede reden om te kiezen voor MI heeft te maken met de toenemende complexiteit van methoden die de ontbrekende data meenemen in het schattingsproces wanneer het modelleren van de data een grotere uitdaging vormt. Men kan bijvoorbeeld denken aan grote hoeveelheden variabelen of complexe univariate of zelfs multivariate verdelingen. In het geval van MI hebben dergelijke ingewikkeldheden voornamelijk betrekking op het imputatiestadium en blijft de analyse relatief eenvoudig. Met andere woorden: zodra plausibele invullingen verkregen zijn is het niet al te moeilijk om een antwoord te vinden op de onderzoeksvraag.

Dit proefschrift richt zich op het vinden van invullingen die plausibel kunnen worden geacht. Plausibele imputaties zijn invullingen die echte waarden zouden kunnen zijn geweest wanneer ze wél waren geobserveerd. Deze definitie van plausibiliteit richt zich op de positie van de geïmputeerde waarden, gegeven de rest van de data. In reguliere datasets houdt dit in dat invullingen moeten passen binnen de incomplete variabele (de kolom) en de overige metingen voor de respondent (de rij). Met andere woorden: plausibiliteit omvat niet alleen de geïmputeerde waarde, maar ook het verband van de imputatie met andere (geobserveerde en geïmputeerde waarden in de data. Een eenvoudig voorbeeld vindt men in variabelen die gezamenlijk optellen tot een totaal. Enkel die imputaties die de somstructuur intact laten kunnen als plausibel worden beschouwd.

Het vinden van plausibele imputaties wanneer de data aan bepaalde restricties onderhevig zijn is lastig. De restricties die gelden voor de data zijn immers ook van toepassing op het model dat wordt gebruikt om de imputaties te genereren. De huidige imputatie methoden leiden niet tot imputaties die naar tevredenheid plausibel kunnen worden geacht. In dit proefschrift worden imputatiemethoden voorgesteld die leiden tot plausibele invullingen voor situaties waarin de huidige imputatiemethoden te kort schieten. De behandelde onderzoekssituaties en restricties komen veelvuldig voor binnen onderzoeksdomeinen waarin de statistiek een prominente rol inneemt, zoals officile statistiek, sociale wetenschappen, geologie en de geneeskunde.

In hoofdstuk 2 wordt gedemonstreerd hoe op een efficiënte wijze plausibele imputaties kunnen worden verkregen voor verdelingen waarvan een groot gedeelte van

de observaties één waarde aanneemt (meestal nul), maar de overige waarden continue zijn verdeeld. Huidige methoden voor het imputeren van dergelijke variabelen bevatten meerdere stappen, dikwijls afhankelijk van vooraf getransformeerde data en zijn in veel gevallen verminderd efficient. De voorgestelde methode bevat slechts één enkele stap, waarbij er geen noodzaak is tot het transformeren van de data om tot plausibele imputaties en valide gevolgtrekkingen te komen.

Hoofdstuk 3 laat zien hoe incomplete multilevel data op een plausibele wijze kunnen worden geïmputeerd. Multilevel data hebben als eigenschap dat respondenten gezamenlijke karakteristieken delen en op basis van deze karakteristieken gegroepeerd kunnen worden in clusters (ook wel: klassen). Indien de clusterstructuur tijdens het imputatieproces wordt genegeerd, zullen de geïmputeerde waarden niet voldoen aan de structuur die op de geobserveerde data van toepassing is. De structuur zal hierdoor afzwakken en de analyses op basis van de gecompleteerde data kunnen sterk vertekende resultaten laten zien. Er wordt gedemonstreerd hoe men op een relatief eenvoudige wijze de clusterstructuur kan meenemen in het imputatieproces. Ook wordt er in Hoofdstuk 3 een vergelijking gemaakt tussen de voorgestelde methode en reeds bestaande imputatiemethoden, waaronder methoden die specifiek zijn ontwikkeld voor het imputeren van multilevel data.

Veel toegepaste onderzoekers gebruiken gekwadrateerde termen in hun analysemodellen. Het is algemeen bekend dat het model dat gebruikt wordt voor het verkrijgen van de imputaties tenminste alle relaties moet omvatten die van wetenschappelijk belang zijn. Dit houdt in dat voor iedere gekwadrateerde term, zowel het kwadraat als de originele variabele in het imputatiemodel meegenomen moeten worden. Bij het genereren van plausibele imputaties dient vervolgens het verband tussen de originele variabele en het kwadraat van deze variabele bewaard te blijven. Een gekwadrateerde waarde die geen enkele relatie heeft tot haar wortel kan immers niet plausibel worden geacht. In Hoofdstuk 4 wordt uiteengezet hoe plausibele imputaties kunnen worden verkregen wanneer gekwadrateerde termen onderdeel uitmaken van het imputatiemodel.

Binnen veel domeinen van wetenschap wordt data verzameld die onderhevig zijn aan een bepaalde compositionele structuur. Compositionele data kunnen worden omschreven als een set van delen die optellen naar een totaal. Het imputeren van compositionele data vormt een uitdaging omdat de invullingen dienen te gehoorzamen aan de vereiste structuur, maar strikt niet-negatief dienen te zijn. Hoofdstuk 5 introduceert een imputatieaanpak die om kan gaan met ingewikkeld gelaagde compositionele structuren en resulteert in plausibele imputaties die vasthouden aan de structuur van de data.

Het laatste hoofdstuk behandelt een nieuwe aanpak voor het samenvoegen van de verschillende analyses to één enkele gevolgtrekking. Deze aanpak is in het bijzonder aantrekkelijk voor onderzoeksdoeleinden waarin computersimulaties worden gedraaid. In simulatieonderzoek wordt doorgaans een steekproef getrokken uit een theoretische verdeling die dienst doet als de populatie. Wanneer er geen theoretische verdeling mogelijk is, kiest men voor een 'design-based' simulatieaanpak waarin een steekproef wordt getrokken uit een echte, geobserveerde dataset die groot genoeg wordt geacht.

Beide simulatieaanpakken introduceren steekproefvariantie in het onderzoek. Deze vorm van variantie is echter niet specifiek interessant wanneer het evalueren van imputatiemethoden het doel van de simulatiestudie is. Hoofdstuk 6 demonstreert daarom een vereenvoudiging van de gebruikelijke regels voor het samenvoegen van analyseresultaten voor situaties waarin de steekproefvariantie niet interessant is. Er wordt aangetoond dat de vereenvoudigde regels ook dienen te worden gebruikt in situaties waarin de grootte van de populatie een restrictie vormt en in essentie alle eenheden in de populatie zijn geobserveerd.