Statistica Neerlandica (2014) Vol. 68, nr. 1, pp. 61–90
doi:10.1111/stan.12023

Predictive mean matching imputation of semicontinuous variables

Gerko Vink*

Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands and Division of Methodology and Quality, Statistics Netherlands, The Hague, the Netherlands

Laurence E. Frank

Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands

Jeroen Pannekoek

Division of Methodology and Quality, Statistics Netherlands, The Hague, the Netherlands

Stef van Buuren

Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands and Netherlands Organization for Applied Scientific Research TNO, Leiden, the Netherlands

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.

Keywords and Phrases: multiple imputation, point mass, predictive mean matching, semicontinuous data, skewed data.

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

*g.vink@uu.nl

Published by Wiley Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

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).

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 (Rubin, 1987; Schafer, 1997; Little and Rubin, 2002). Univariate models for modeling semicontinuous data have been developed as well as the Tobit model (Tobin, 1958; Amemiya, 1984) 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, Royston and Wood, 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, Burton and Rivero-Arias (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, and IVEware` (Raghunathan, Solenberger and Van Hoewyk, 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, Little and Raghunathan, 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) focused on readily available software implementations, setting aside methods specifically designed for semicontinuously distributed data (Schafer and Olsen, 1999; Olsen and Schafer, 2001; Javaras and Van Dyk, 2003). 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, Kowarik and Filzmoser, 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).

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 Imputation methods

2.1 Notation and preliminaries

Let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ 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, $X = (X_1, \ldots, X_j)$ is a set of j fully observed covariates, where X_{obs} and X_{mis} correspond to the observed 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. We must note that we limit our research to the univariate case, although problems could be approached iteratively to solve the multivariate problem, without loss of generality.

To impute missing values in Y and to asses variances and confidence intervals for estimators based on the imputed data, we use multiple imputation `mi` 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 square 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 $N(\hat{\beta}, V(\hat{\beta}))$, that is, normal with mean $\hat{\beta}$ and covariance matrix $V(\hat{\beta}) = \sigma^2 \left(X_{\text{obs}}^T X_{\text{obs}}\right)^{-1}$, and the posterior distribution for σ^2 is given by $\hat{\epsilon}^T \hat{\epsilon}/A$, with A a chi-square variate with $n_{\text{obs}} - r$ 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 $Y_{\text{mis},i}$ from $N\left(X_i^T \beta^*, \sigma^{2*}\right)$.

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: $\pi_i = \text{expit}\left(X_i^T \gamma\right) = \exp\left(X_i^T \gamma\right) / \left[1 + \exp\left(X_i^T \gamma\right)\right]$. Using an uninformative prior for γ, the corresponding posterior distribution is approximately $\mathrm{N}(\hat{\gamma}, \hat{V}(\hat{\gamma}))$ 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 $W_{\mathrm{mis},i}$ 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 $\pi^* = \mathrm{expit}(X_i^T \gamma^*)$.

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 $n_{\mathrm{obs}} - r$ degrees of freedom.
3. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}\left(X_{\mathrm{obs}}^T X_{\mathrm{obs}}\right)^{-1}$.
4. Calculate $\hat{Y}_{\mathrm{obs}} = X_{\mathrm{obs}}\,\hat{\beta}$ and $\hat{Y}_{\mathrm{mis}} = X_{\mathrm{mis}}\,\beta^*$.
5. For each $\hat{Y}_{\mathrm{mis},i}$, find $\Delta = |\hat{Y}_{\mathrm{obs}} - \hat{Y}_{\mathrm{mis},i}|$.
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 $Y_{\mathrm{obs},i}$ as the imputation.
7. Repeat steps 1–6 m times, each time saving the completed dataset.

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 $\Delta = \min|\hat{Y}_{\mathrm{obs}} - \hat{Y}_{\mathrm{mis},i}|$ and selects the corresponding $Y_{\mathrm{obs},i}$ as the imputation.

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

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

$$Z_i = \begin{cases} g(Y_i) & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0 \end{cases}, \tag{2}$$

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 $\pi_i^* = \text{expit}\left(X_i^T\gamma^*\right)$ 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 $i \in \text{obs}$.
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 $n_{\text{obs}.1} - r$ degrees of freedom, with $n_{\text{obs}.1}$ the number of observed elements given $W_i = 1$.
6. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}\left(X_{\text{obs}}^T X_{\text{obs}}\right)^{-1}$.
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 dataset. Note that steps 1 and 4 do not change and need to be performed only once. Further, steps 4–7 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.

2.4 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 expectation–maximization 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.

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, Templ and Filzmoser (2010a, 2010b)] and is often used in the case of performance assessment of imputation procedures in this field.

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 $Q \sim \mathrm{N}(5,1)$ to which we assign a point mass at zero by drawing from a binomial distribution with a 30 (population 1) or 50 (population 2) percent 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

$$R_{QX} = \begin{bmatrix} Q & X_1 & X_2 & X_3 & X_4 \\ 1 & & & & \\ 0.80 & 1 & & & \\ 0.50 & 0.4 & 1 & & \\ 0.30 & 0.24 & 0.15 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

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

$$\mathrm{SDS}_{X_{ij}} = \mathrm{SDS}_{Q_i} * \rho_{QX_j} + \epsilon_i,$$

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 $\mathrm{N}(0, \sqrt{1 - \rho^2_{YX_j}})$.

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 1 displays histograms for Y_1 through Y_5 with a 50% point mass at zero.

Combining the set of transformed variables $Y = (Y_1, \ldots, Y_5)$ with the variables in $X = (X_1, \ldots, X_4)$ 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

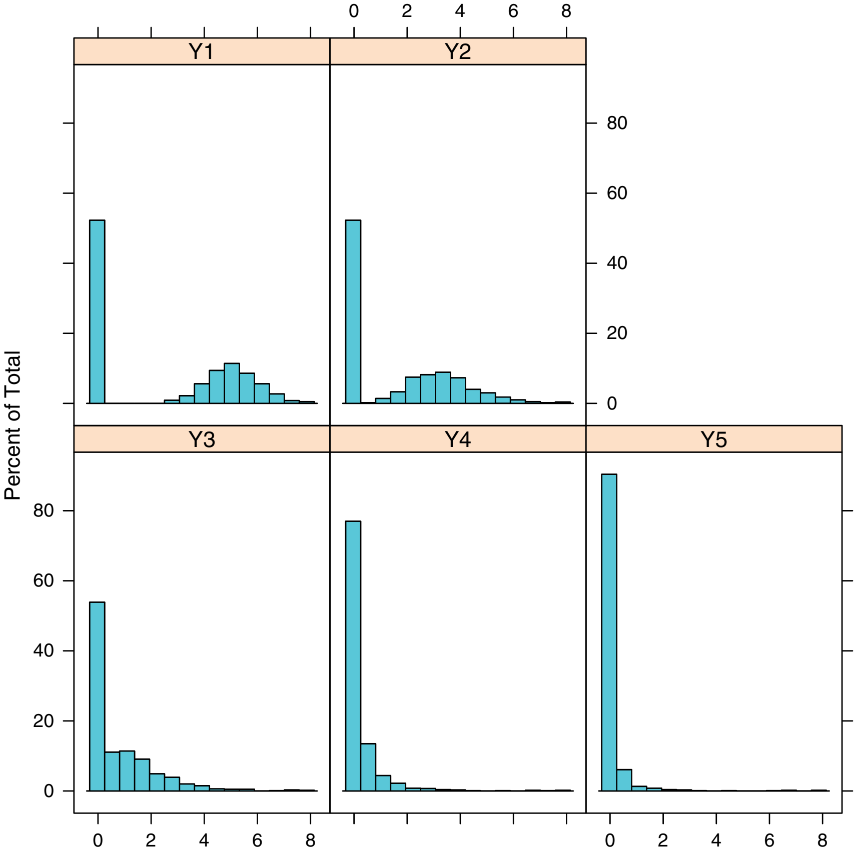

Fig. 1. Generated semicontinuous variables ($\mathbf{Y_1} - \mathbf{Y_5}$) with a point mass at 50%.

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. 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).

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.

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:

$$P\left(R=0|Y_{\text{obs}},Y_{\text{mis}},X_j\right)=P\left(R=0|Y_{\text{obs}},X_j\right)$$

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, $a=\left(-\overline{X}_j+X_{ij}\right)/\sigma_{X_j}$ 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 $a=\left(\overline{X}_j-X_{ij}\right)/\sigma_{X_j}$. Choosing $a=0.75-\left[\left(\overline{X}_j-X_{ij}\right)/\sigma_{X_j}\right]$, or $a=-0.75+\left[\left(\overline{X}_j-X_{ij}\right)/\sigma_{X_j}\right]$, 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 performed 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 postprocessing 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 preprocessed 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.

3.4 Evaluation of imputations

In the case of a simulated dataset, evaluations can be performed because 'truth' is known. In case of a real-life dataset, containing observed missingness, this cannot be performed, 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, Gelman and Levy (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 assess 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.

4 Univariate results

4.1 Bias of the mean

Tables 1 and 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.

Table 1. Univariate simulation results for X_1 over 100 simulations

			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

Notes: 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 $\mathbf{Y_1}$ through $\mathbf{Y_3}$. All cases represent a sample size of $n = 500$ and 50% MAR missingness.

Table 2. Univariate simulation results for X_4 over 100 simulations

			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

Notes: 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 $\mathbf{Y_1}$ through $\mathbf{Y_3}$. All cases represent a sample size of $n = 500$ and 50% MAR missingness.

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 (Figure 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).

4.2 Bias of the correlation with the covariate

Figure 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

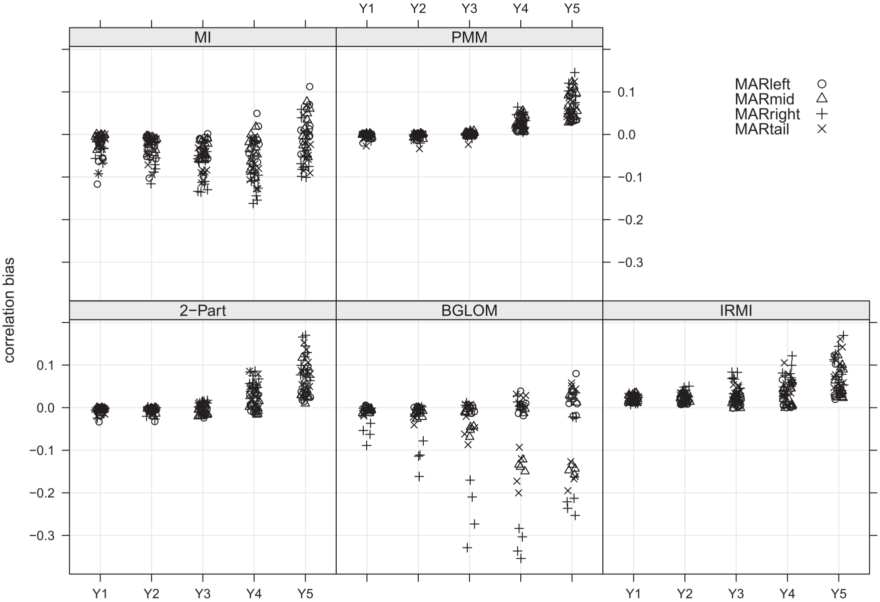

Fig. 2. Bias of the correlation with the covariate X_1 for different imputation methods over 100 simulations.

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

4.3 Bias of the median

Estimating the median for Y_1 through Y_5 from imputed data can lean 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 (Figure 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.

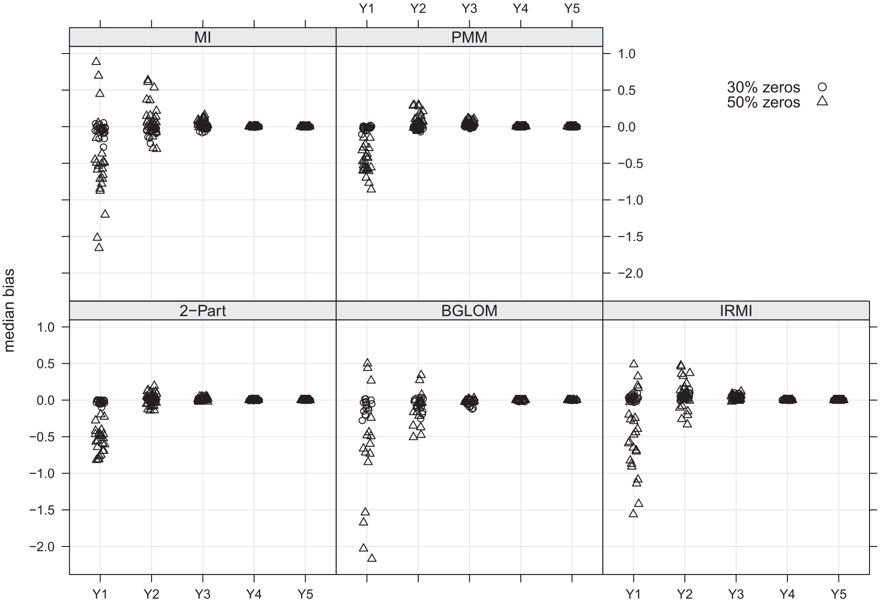

Fig. 3. Bias of the median for different sizes of the point mass over 100 simulations given covariate X_1.

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 (Figure 4). However, we can see in Figure 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 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.

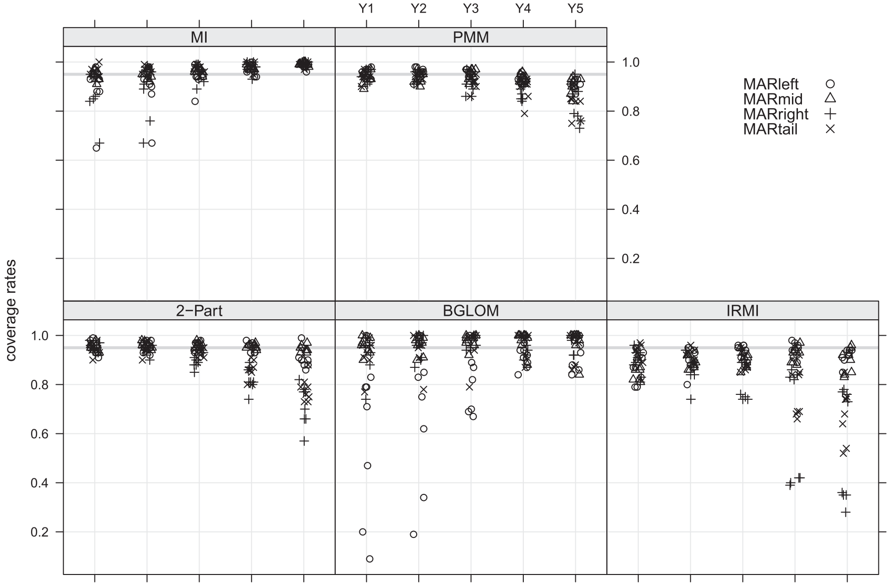

Fig. 4. Coverage rates for different imputation methods over 100 simulations using covariate X_1.

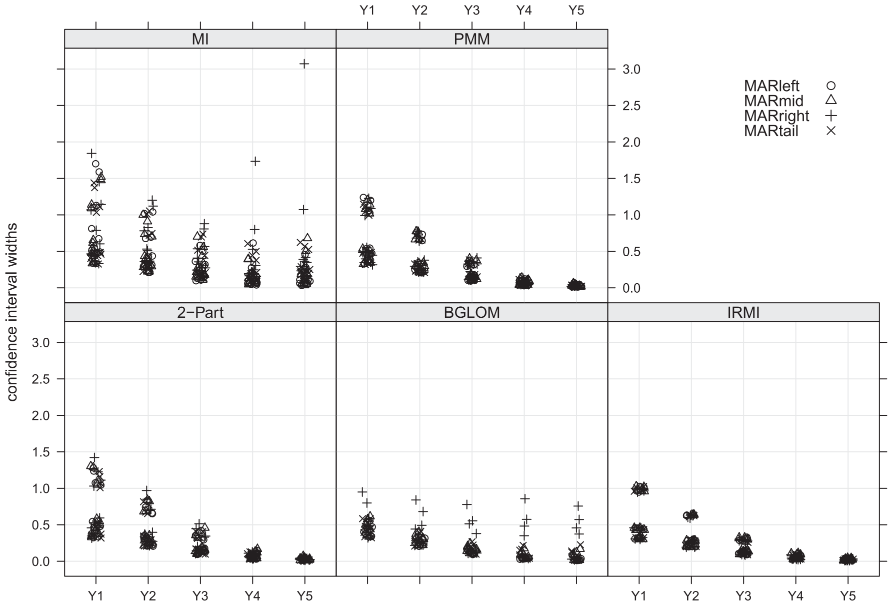

Fig. 5. Confidence interval widths for different imputation methods over 100 simulations using covariate X_1.

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.

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.

4.5 Point mass

Tables 1 and 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 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.

The point mass estimated by mi is acceptable in the case of a high-correlating covariate, although PMM, irmi, and the two-part 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 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 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 2.2.

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

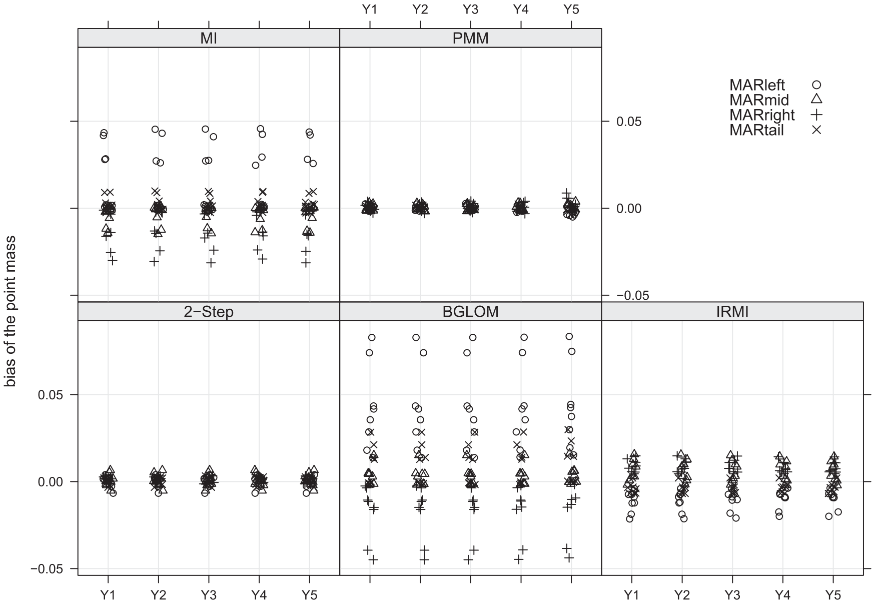

Fig. 6. Bias of the estimated size of the point mass for different imputation methods over 100 simulations using covariate X_1.

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.

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, two-part imputation and `irmi`, it holds that when skewness increases, these model-based methods tend to represent a normal curve again (Figure 7).

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 resembling the original distribution closer. 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.

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.

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

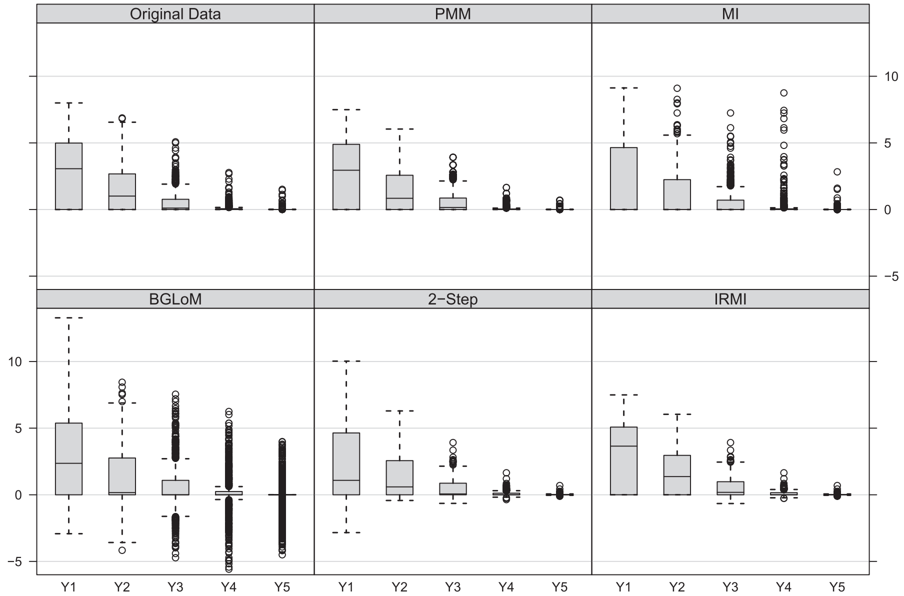

Fig. 7. Right-tailed MAR missingness: boxplots of the original data and imputed data for five imputation methods for 50% missing data. Imputations are based on covariate X_1.

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

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

We create multivariate missingness following the procedure as described in Section 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.

6 Multivariate results

6.1 Multivariate normal

The results of the multivariate normal simulations can be found in Table 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 `mi` 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.

Table 3. Normal simulations: biases and coverage rates for the mean of the multivariate normal simulation

		A	B	C	D	Y_1	cov Y_1	Y_2	cov Y_2	ρ_{Y_1,Y_2}
CCA	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
	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
PMM	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
	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
2-Part	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
	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
MI	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
	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
IRMI	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
	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
BGLoM	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
	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

Notes: All biase Y_1s depict the average simulation value subtracted by the population value. Please note that the biases in A, B, C, and D are completed data proportions minus true proportions.

*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.

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 therefore still useful. The results of the multivariate simulation with skewed target variables can be found in Table 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.

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 (cf. Sections 4.4 and 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.

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

Table 4. Skewed simulations: biases and coverage rates for the mean of the multivariate skewed simulation

		A	B	C	D		cov Y_1	Y_2	cov Y_2	ρ_{Y_1,Y_2}
CCA	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
	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
PMM	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
	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
2-Part	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
	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
MI	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
	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
IRMI	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
	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
BGLoM	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
	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

Notes: All biases depict the average simulation value subtracted by the population value. Please note that the biases in A, B, C, and D are completed data proportions minus true proportions.

semicontinuous data. The results of the multivariate simulation with skewed target variables with outliers can be found in Table 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.

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.

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.

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

Table 5. Outlier simulation: biases and coverage rates for the mean of the multivariate skewed simulation with outliers

		A	B	C	D	Y_1	cov Y_1	Y_2	cov 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

Notes: All biases depict the average simulation value (with outliers) subtracted by the population value (without outliers). Please note that the biases in A, B, C, and D are completed data proportions minus true proportions.

*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.

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

7.2 Dutch Wholesalers 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 3.3) on TEMPS with the total amount of employees (EMPL) as a covariate. Left-tailed

Table 6. Comparison between true and imputed ITSR for all imputation methods

	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

Notes: Depicted are the total amount of zeros, the amount of values in cells A, B, C, and D, the correlation $\boldsymbol{\rho}_\mathbf{D}$ of values in cell D, the total correlation $\boldsymbol{\rho}$, mean ITSR after imputation and the width of the confidence interval.

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 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 therefore underestimates the mean and sum of TEMPS, but correlation bias is lowest of all investigated methods.

8 Conclusions

How does PMM compare to specialized methods such as `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

Table 7. Comparison between true and imputed TEMPS for all imputation methods

	Zero	ρ	Mean	ciw	Sum
TEMPS	304.00	0.48	5.02	—	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

Notes: Depicted are the total amount of zeros, the correlation between TEMPS and EMPL $\boldsymbol{\rho}$, mean TEMPS after imputation, and the width of the confidence interval.

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 is 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, as 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 performed 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 types 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, PMM is at least as good for imputing semicontinuous data than dedicated methods for such data. PMM is very flexible as a method, because of 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.

References

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

ALFONS, A., M. TEMPL and P. FILZMOSER (2010a), Applications of statistical simulation in the case of eu-silc: using the r package simframe, *Journal of Statistical Software* **37**, 17.

ALFONS, A., M. TEMPL and P. FILZMOSER (2010b), An object-oriented framework for statistical simulation: the r package simframe, *Journal of Statistical Software* **37**, 1–36.

AMEMIYA, T. (1984), Tobit models: a survey, *Journal of Econometrics* **24**, 3–61.

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

DUAN, N., W. G., MANNING JR., C. N., MORRIS and J. P. NEWHOUSE (1983), A comparison of alternative models for the demand for medical care, *Journal of Business & Economic Statistics* **1**, 115–126.

HECKMAN, J. (1974), Shadow prices, market wages, and labor supply, *Econometrica* **42**, 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., R. LITTLE and T. RAGHUNATHAN (2002), Survey nonresponse, chapter *Multivariate imputation of coarsened survey data on household wealth*, pages 357–371. Wiley.

JAVARAS, K. N. and D. A. VAN DYK (2003), Multiple imputation for incomplete data with semicontinuous variables, *Journal of the American Statistical Association* **98**, 703–715.

LITTLE, R. (1988), Missing-data adjustments in large surveys, *Journal of Business and Economic Statistics* **6**, 287–296.

LITTLE, R. J. A. and D. B. RUBIN (2002), *Statistical analysis with missing data*. Wiley-Interscience, New York.

MANNING, W., C. MORRIS, J. NEWHOUSE, L. ORR, N. DUAN, E. KEELER, A. LEIBOWITZ, K. MARQUIS, M. MARQUIS and C. PHELPS (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.*

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

OLSEN, M. K. and J. L. SCHAFER (2001), A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.

RAGHUNATHAN, T., P. SOLENBERGER and J. VAN HOEWYK (2002), IVEware: imputation and variance estimation software. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

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.

Sargasso.nl (2012), De haagse twitter stolp.

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

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

SU, Y., A. GELMAN, J. HILL and M. YAJIMA (2011), Multiple imputation with diagnostics (mi) in r: opening windows into the black box, *Journal of Statistical Software* **45**, 1–31.

TEMPL, M., A. KOWARIK and P. FILZMOSER (2011), Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics and Data Analysis* **55**, 2793–2806.

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

VAN BUUREN, S. (2012), *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton, FL.

VAN BUUREN, S. and C. GROOTHUIS-OUDSHOORN (2011), MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**, 1–67.

WHITE, I., P. ROYSTON and A. WOOD (2011), Multiple imputation using chained equations: issues and guidance for practice, *Statistics in Medicine* **30**, 377–399.

YU, L., A. BURTON and O. RIVERO-ARIAS (2007), Evaluation of software for multiple imputation of semicontinuous data. *Statistical Methods in Medical Research*, **16**, 243–258.

Received: 15 January 2014.