Generalizing Univariate Predictive Mean Matching to Impute Multiple Variables Simultaneously

Mingyang Cai(✉), Stef van Buuren, and Gerko Vink

Utrecht University, Padualaan 14, 3584 CH Utrecht, Netherlands
m.cai@uu.nl

Abstract. Predictive mean matching (PMM) is an easy-to-use and versatile univariate imputation approach. It is robust against transformations of the incomplete variable and violation of the normal model. However, univariate imputation methods cannot directly preserve multivariate relations in the imputed data. We wish to extend PMM to a multivariate method to produce imputations that are consistent with the knowledge of derived data (e.g., data transformations, interactions, sum restrictions, range restrictions, and polynomials). This paper proposes multivariate predictive mean matching (MPMM), which can impute incomplete variables simultaneously. Instead of the normal linear model, we apply canonical regression analysis to calculate the predicted value used for donor selection. To evaluate the performance of MPMM, we compared it with other imputation approaches under four scenarios: 1) multivariate normal distributed data, 2) linear regression with quadratic terms; 3) linear regression with interaction terms; 4) incomplete data with inequality restrictions. The simulation study shows that with moderate missingness patterns, MPMM provides plausible imputations at the univariate level and preserves relations in the data.

Keywords: Missing data · Multiple imputation · Block imputation · Predictive mean matching · Multivariate analysis · Canonical regression analysis

1 Introduction

Multiple imputation (MI) is a popular statistical method for the analysis of missing data problems. To provide valid inferences from the incomplete data, the analysis procedure of MI consists of three steps. First, in the imputation step, missing values are drawn from a plausible distribution (e.g., posterior distributions for Bayesian model-based approaches and a cluster of candidate donors for non-parametric approaches) to generate several (m) complete datasets. The value of m commonly varies between 3 to 10. Second, in the analysis step, complete data analysis are used to estimate the quantity of scientific interest for each imputed data set. This step yields m separate analyses because imputed

K. Arai (Ed.): SAI 2022, LNNS 506, pp. 75–91, 2022.
https://doi.org/10.1007/978-3-031-10461-9_5

datasets are different. Finally, in the pooling step, m results are aggregated into a single result by Rubin's rules, accounting for the uncertainty of estimates due to the missing data [1].

Two widely used strategies for imputing multivariate missing data are joint modeling (JM) and fully conditional specification (FCS). Joint modeling was proposed by Rubin [1] and especially developed by Shafer [2]. Given that the data is assumed to follow a multivariate distribution, all incomplete variables are generally imputed by drawing from the joint posterior predictive distribution conditional on other variables. Fully conditional specification, which was developed by Van Buuren [3], follows an iterative scheme that imputes each incomplete variable based on a conditionally specified model [3]. Fully conditional specification allows for tremendous flexibility in multivariate model design and flexibility in imputing non-normal variables, especially discrete variables [4]. However, FCS may suffer from incompatibility problems, and computational shortcuts like the sweep operator cannot be applied to facilitate computation [5]. On the other hand, joint modeling possesses more solid theoretical guarantees. With increasing incomplete variables, JM may lead to unrealistically large models and a lack of flexibility, which will not occur under FCS.

In practice, there are often extra structures in the missing data which are not modelled properly. Suppose there are two jointly missing variables $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. There may be restrictions on the sum of $\boldsymbol{X_1}$ and $\boldsymbol{X_1}$ (e.g., $\boldsymbol{X_1} + \boldsymbol{X_1} = C$, where C is a fixed value) and the rank of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ (e.g., $\boldsymbol{X_1} > \boldsymbol{X_2}$), data transformations (e.g., $\boldsymbol{X_2} = log(\boldsymbol{X_1})$, $\boldsymbol{X_2} = \boldsymbol{X_1}^2$) or interaction terms included in the data ($\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ are jointly missing, where $\boldsymbol{X_3} = \boldsymbol{X_1} * \boldsymbol{X_2}$). In this paper, we would focus on the setting of structures between two jointly missing variables, which is a simple scenario to illustrate.

The two popular approaches of MI mentioned before may not be appropriate for modeling the relations among multiple variables in the missing data. Joint modeling may lack the flexibility of modeling the relations explicitly, and FCS imputes each missing variable separately, which may not ensure that the imputation remains consistent with the observed relations among multiple variables.

Van Buuren [5] suggested block imputation, which combines the strong points of joint modeling and fully conditional specification. The general idea is to place incomplete variables into blocks and apply multivariate imputation methods to the block. Joint modeling can be viewed as a "single block" imputation method. In contrast, FCS is strictly a multiple blocks imputation method, where the number of blocks equals the number of incomplete columns in the data. It is feasible to consider the relations among a set of missing variables if we specify them as a single block and perform the MI iteratively over the blocks.

Based on the rationale of block imputation, we extend univariate predictive mean matching to the multivariate case to allow for the joint imputation of blocks of variables. The general idea is to match the incomplete case to one of the complete cases by applying canonical regression analysis and imputing the variables in a block entirely from the matched case [6]. We shall refer to the multivariate extension of PMM as *multivariate predictive mean matching* (MPMM).

Predictive mean matching (PMM) is a user-friendly and versatile non-parametric imputation method. Multiple imputation by chained equation (`MICE`), which is a popular software package in `R` for imputing incomplete multivariate data by Fully Conditional Specification (FCS), sets the PMM as the default imputation approach [7]. We tailor PMM to the block imputation framework, which will widen its application. More computational details and properties of PMM would be addressed in Sect. 2.

For a comprehensive overview of missing data analysis, we refer to Little and Rubin [8] for a comparison of approaches to missing data other than multiple imputation (e.g., ad-hoc methods, maximum likelihood estimation and weighting methods). Schafer [9], Sinharay et al. [10] and Allison [11] introduced basic concepts and general methods of MI. Schafer and Graham [12] discussed practical issues of application of MI. Various sophisticated missing data analysis were developed on the fields of multilevel model [13], structural equation modeling [14,15], longitudinal data analysis [16,17] and meta-analysis [18]. Schafer [19] compared Bayesian MI methods with maximum likelihood estimation. Seaman and White [20] gave an overview of the use of inverse probability weighting in missing data problems. Ibrahim et al. [21] provided a review of various advanced missing data methods. Because an increasing number of missing data methodologies emerged, MI as well as other approaches were applied in many fields (e.g., epidemiology, psychology and sociology) and implemented in many statistical software packages (e.g., `mice` and `mi` in `R`, `IVEWARE` in `SAS`, `ice` in `STATA` and module `MVA` in `SPSS`) [7].

The following section will outline canonical regression analysis, introduce predictive mean matching (PMM), and connect the techniques to propose multivariate predictive mean matching (MPMM). Section 3 provides a simple comparison between PMM and MPMM. Section 4 is a simulation study investigating whether MPMM yields valid estimates and preserves functional relations between imputed values. The discussion closes the paper.

2 Multivariate Predictive Mean Matching

2.1 Canonical Regression Analysis (CRA)

Canonical regression analysis is a derivation and an asymmetric version of canonical correlation analysis (CCA). It aims to look for a linear combination of covariates that predicts a linear combination of outcomes optimally in a least-squares sense [22]. The basic idea of canonical regression analysis is quite old and has been discussed under different names, such as Rank-reduced regression [23] and partial least squares [24].

Let us consider the equation

$$\boldsymbol{\alpha'Y} = \boldsymbol{\beta X} + \epsilon. \tag{1}$$

We aim to minimize the variance ϵ with respect to α and β under some restrictions. CRA can be implemented by maximizing the squared multiple correlation coefficient for the regression of $\alpha'Y$ on X, which can be written as

$$R^2_{\alpha' y.x} = \frac{\boldsymbol{\alpha' \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \alpha}}{\boldsymbol{\alpha' \Sigma_{yy} \alpha}}, \tag{2}$$

where $R^2_{\alpha' y.x}$ is the ratio of the amount of variance of $\boldsymbol{\alpha' Y}$ accounted for by the covariates $\boldsymbol{X}$ to the total variance. According to McDonald [25], maximization of the above equation leads to eigenvalue decomposition. The solution is that $\boldsymbol{\alpha}$ is the right-hand eigenvector of $\boldsymbol{\Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}$ corresponding to its greatest eigenvalue. After reducing the rank of $\boldsymbol{\alpha' Y}$ to 1, we could estimate $\boldsymbol{\beta}$ by multivariate regression analysis.

2.2 Predictive Mean Matching (PMM)

PMM was first proposed by Rubin [26]and formalized by Little [6]. It can be viewed as an extension of the k nearest neighbor method. PMM calculates the estimated value of the missing variable through a specified imputation model (e.g., linear imputation model). The method selects a set of candidate donors (typically, the number of candidate donors is 5) from all complete cases whose estimated values are closest to the estimated value of the missing unit. The unobserved value is imputed by randomly drawing one of the observed values of the candidate donors [5].

Computational Details. We elaborate the algorithm of predictive mean matching for the clear illustration of its merger with canonical regression analysis [27]. $\boldsymbol{X_{obs}}$, a $N_{obs} \times j$ matrix, denotes the observed part of predictors and $\boldsymbol{X_{mis}}$, a $N_{mis} \times j$ matrix, denotes the missing part of predictors.

1. Use linear regression of $\boldsymbol{Y_{obs}}$ given $\boldsymbol{X_{obs}}$ to estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\epsilon}$ through ordinary least squares
2. Draw $\sigma^{2*} = \hat{\epsilon}^{\mathsf{T}} \hat{\epsilon} / A$, where A is a χ^2 variate with $N_{obs} - j$ degrees of freedom
3. Draw $\boldsymbol{\beta^*}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{\sigma^{2*}(X_{obs}^{\mathsf{T}} X_{obs})^{-1}}$
4. Calculate $\boldsymbol{\hat{V}_{obs} = X_{obs} \hat{\beta}}$ and $\boldsymbol{\hat{V}_{mis} = X_{mis} \beta^*}$
5. For each missing cell $y_{mis,n}$, where $n = 1, \cdots, N_{mis}$
   (a) Find $\Delta = |\hat{v}_{mis,n} - \hat{v}_{obs,k}|$ for all $k = 1, \cdots, N_{obs}$
   (b) Pick several observed entries y_{obs}, 5 as default in *mice.impute.pmm*, with the smallest distance defined in step 5(a)
   (c) Randomly draw one of the y_{obs} which are picked in the previous step to impute $y_{mis,n}$
6. Repeat steps 1–5 m times and save m completed datasets.

Predictive mean matching has been proven to perform well in a wide range of simulation studies and is an attractive way to impute missing data [7,27–30]. More precisely, PMM has the appealing features that the imputed values 1) follow the potential distributions of the data and 2) are always within the range of observed data because imputed values are replaced by real observed values [5]. For the same reason, PMM yields acceptable imputations even when normality

assumptions are violated [30]. In cases where the observed values follow a skewed distribution, the imputations will also be skewed. If observations are strictly positive, so will the imputations from PMM be. Furthermore, since PMM does not rely on model assumptions, it alleviates the adverse impact when the imputation model is misspecified [31].

Although PMM was developed for situations with only a single incomplete variable, it is easy to implement it under a fully conditionally specification framework for imputing multivariate missing data. However, the application of PMM under FCS framework is only limited to univariate imputation. Therefore, it may distort the multivariate relations in the imputations and narrow the application of the method to more complex data structures. For example, Seaman et al. [32] concluded that a univariate implementation of predictive mean matching is not advised to produce plausible estimates when the analysis model contains non-linear terms. As a multivariate extension to PMM, we expect that MPMM could yield plausible and consistent imputations when missing covariates include polynomial or interaction terms.

2.3 Multivariate Predictive Mean Matching (MPMM)

For illustration, we present the algorithm with one missing data pattern. The appendix discusses the extension to cases with multiple missing patterns. Let $\boldsymbol{Y} = (\boldsymbol{Y_1}, \cdots, \boldsymbol{Y_I})$ and $\boldsymbol{X} = (\boldsymbol{X_1}, \cdots, \boldsymbol{X_J})$ be two sets of I jointly incomplete variables and J complete quantitative variables, respectively. Let $\boldsymbol{V} = \boldsymbol{\alpha Y}$ denotes the linear combination of multiple response variables and $\boldsymbol{X}$ denotes predictors with j dimensions.

1. Use the observed data to estimate the $(I + J) \times (I + J)$ covariance matrix
$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{y_{obs}y_{obs}}} & \boldsymbol{\Sigma}_{\boldsymbol{y_{obs}x_{obs}}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x_{obs}y_{obs}}} & \boldsymbol{\Sigma}_{\boldsymbol{x_{obs}x_{obs}}} \end{pmatrix}$$
2. Find the largest eigenvalue λ^2 of $\boldsymbol{\Sigma}^{\boldsymbol{-1}}_{\boldsymbol{y_{obs}y_{obs}}} \boldsymbol{\Sigma}_{\boldsymbol{y_{obs}x_{obs}}} \boldsymbol{\Sigma}^{\boldsymbol{-1}}_{\boldsymbol{x_{obs}x_{obs}}} \boldsymbol{\Sigma}_{\boldsymbol{x_{obs}y_{obs}}}$ and its corresponding right-hand eigenvector α
3. Calculate the linear combination $\boldsymbol{\alpha' Y}$ for all completely observed individuals in the sample: $\boldsymbol{V_{obs} = \alpha' Y_{obs}}$
4. Use linear regression of $\boldsymbol{V_{obs}}$ given $\boldsymbol{X_{obs}}$ to estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\epsilon}$ through ordinary least squares
5. Draw $\sigma^{2*} = \hat{\epsilon}^{\mathsf{T}}\hat{\epsilon}/A$, where A is a χ^2 variate with $N_{obs} - j$ degrees of freedom
6. Draw β^* from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$and covariance matrix $\boldsymbol{\sigma^{2*}(X^{\mathsf{T}}_{obs} X_{obs})^{-1}}$
7. Calculate $\hat{\boldsymbol{V}}_{\boldsymbol{obs}} = \boldsymbol{X_{obs}}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{V}}_{\boldsymbol{mis}} = \boldsymbol{X_{mis}\beta^*}$
8. For each missing vector $\boldsymbol{y_{mis,n}}$, where $n = 1, \cdots, N_{mis}$
   (a) Find $\Delta = |\hat{v}_{mis,n} - \hat{v}_{obs,k}|$ for all $k = 1, \cdots, N_{obs}$
   (b) Pick several observed components $y_{obs} = \{y_{1,obs}, \cdots, y_{I,obs}\}$, 5 as default, with the smallest distance defined in step 8(a)
   (c) Randomly draw one of the y_{obs} which are picked in the previous step to impute $y_{mis,n}$
9. Repeat steps 5–8 m times and save m completed datasets.

We also tried other methods of multivariate analysis, such as multivariate regression analysis (MRA) [33] and redundancy analysis (RA) [34]. However, imputation models specified by MRA or RA are not appropriate because of the assumed independence between missing variables. The violation of this assumption leads to less sensible imputations when there are extra relations among missing covariates.

3 Comparison Between PMM and MPMM

We shall illustrate that although MPMM is a multivariate imputation method, where the whole missing component is assigned entirely from the matching donor, the derived imputed datasets are also plausible at the univariate level.

3.1 Simulation Conditions

The predictors were generated by a multivariate distribution

$$\begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \boldsymbol{X_3} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 12 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix} \right].$$

The responses were generated based on the multivariate linear model

$$\begin{pmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \boldsymbol{Y_3} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 3\boldsymbol{X_1} + \boldsymbol{X_2} + 2\boldsymbol{X_3} \\ \boldsymbol{X_1} + 5\boldsymbol{X_2} + 2\boldsymbol{X_3} \\ 5\boldsymbol{X_1} + 3\boldsymbol{X_2} + \boldsymbol{X_3} \end{pmatrix}, \begin{pmatrix} 4 & 4\rho & 4\rho \\ 4\rho & 4 & 4\rho \\ 4\rho & 4\rho & 4 \end{pmatrix} \right],$$

where ρ denotes the correlation between the predictors $\boldsymbol{X}$. Let $\mathbf{R}$ be the vector of observation indicators whose values are zero if the corresponding variable is missing and one if observed. We simulated missingness such that rows in the set $(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3})$ were always either observed or completely missing. This joint missingness was either completely at random (MCAR) with $P(\mathbf{R} = 0|\mathbf{X}, \mathbf{Y}) = 0.4$ or right-tailed missing at random (MARright) with $P(\mathbf{R} = 0|\mathbf{X}, \mathbf{Y}) = \frac{e^a}{1+e^a}$, where $a = \alpha_0 + \boldsymbol{X_1}/SD(\boldsymbol{X_1})$ and α_0 was chosen to make the probability of jointly missing $\mathbf{Y}$ equal to 0.4. Missing values were induced with the `ampute` function [35] from the package `MICE` [7] in `R` [36]. The correlation ρ was simulated from 0.2, 0.5 or 0.8 corresponding to a weak, moderate and strong dependence between predictors. The sample size was 2000, and 1000 simulations were repeated for different setups.

For reasons of brevity, we focused our evaluation on the expectation of $\boldsymbol{Y_1}$ and the correlation between $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$. We studied the average bias over 1000 simulations with respect to the designed population value and the coverage rate of nominal 95% confidence interval. Within each simulation, we generated five imputed datasets and combined the statistics into a single inference by using Rubin's combination rules [1].

3.2 Results

Table 1. Simulation results for evaluating whether MPMM provide valid imputations at the univariate level.

ρ	scenario	$E(Y_1)$				$\rho(Y_1, Y_2)$			
		PMM		PMM-CRA		PMM		PMM-CRA	
		bias	cov	bias	cov	bias	cov	bias	cov
0	MCAR	0	0.94	0	0.95	0	0.95	0	0.94
	MAR	0	0.93	0	0.94	0	0.96	0	0.94
0.5	MCAR	0	0.95	0	0.93	0	0.95	0	0.95
	MAR	0	0.94	0	0.94	0	0.94	0	0.94
0.8	MCAR	0	0.93	0	0.94	0.01	0.91	0	0.95
	MAR	0	0.93	0	0.93	0.01	0.93	0	0.94

Table 1 shows the simulation results. In general, MPMM yielded no discernible difference with PMM when focusing on the correlation coefficient $\rho(\boldsymbol{Y_1}, \boldsymbol{Y_2})$. Under the MCAR missingness mechanism, both methods yielded unbiased estimates and displayed coverage rates close to the nominal 95%, and even there was 40% missingness in the joint set $(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3})$. It is notable to see that with MAR-right and high correlation between $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$, PMM had a somewhat reduced coverage rate, which suggests that MPMM yielded more robust results against various correlation coefficients. For estimation of the mean value $E(Y_1)$, MPMM performed similarly to PMM. Both methods yielded plausible imputations with various missingness scenarios and different pre-assumed correlation coefficients.

These initial results suggested that multivariate predictive mean matching could be an alternative to predictive mean matching. If PMM yields sensible imputations, so will PMM-CRA.

4 Evaluation

To investigate the performance of MPMM when there are relations in the incomplete data, we performed the following simulation studies carried out in `R` 4.0.5 [36].

4.1 Linear Regression with Squared Term

We first simulated from a linear regression substantive model with a squared term.

Simulation Conditions. The dependent variable $\boldsymbol{Y}$ was generated according to the analysis model

$$\boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X^2} + \epsilon \tag{3}$$

where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, both predictor $\boldsymbol{X}$ and error term ϵ were assumed as standard normal distributions. These coefficients lead to a strong quadratic association between $\boldsymbol{Y}$ and $\boldsymbol{X}$. A large sample size ($n = 5000$) was created. Simulations were repeated 1000 times so that we could achieve more robust and stable analyses. Forty percent of $\boldsymbol{X}$ and $\boldsymbol{X^2}$ were designed to be jointly missing under five various missingness mechanisms: MCAR, MARleft, MARmid, MARtail, and MARright[1], which means no cases with missing values on either X or X^2 for each mechanism. Missing values were again created with the `ampute` function from the package `MICE` in `R`.

Estimation Methods. We compared the performance of MPMM to four other approaches: *'transform, then impute'* (TTI), *'impute, then transform'* (ITT), polynomial combination method (PC) and substantive model compatible FCS (SMC-FCS). *'Impute, then transform'*, also named as passive imputation, excludes $\boldsymbol{X^2}$ during imputation and appends it with the square of $\boldsymbol{X}$ afterwards. *'Transform, then impute'*, also known as just another variable (JAV), treats the squared term as another variable to be imputed. Both aforementioned methods are proposed by von Hippel [37]. We also apply polynomial combination proposed by Vink and Van Buuren [38]. PC imputes the combination of $\boldsymbol{X}$ and $\boldsymbol{X^2}$ by predicted mean matching and then decomposes it by solving a quadratic equation for $\boldsymbol{X}$. The polynomial combination method is implemented by `mice.impute.quadratic` function in the `R` `MICE` package. Finally, SMC-FCS is proposed by Barlett et al. [39]. In general, it imputes the missing variable based on the formula:

$$\begin{aligned} f(\boldsymbol{X_i}|\boldsymbol{X_{-i}}, \boldsymbol{Y}) &= \tfrac{f(\boldsymbol{X_i}, \boldsymbol{X_{-i}}, \boldsymbol{Y})}{f(\boldsymbol{Y}, \boldsymbol{X_{-i}})} \\ &\propto f(\boldsymbol{Y}|\boldsymbol{X_i}, \boldsymbol{X_{-i}}) f(\boldsymbol{X_i}|\boldsymbol{X_{-i}}). \end{aligned} \tag{4}$$

Provided the scientific model is known and the imputation model is specified precisely (i.e., $f(\boldsymbol{Y}|\boldsymbol{X_i}$ fits the substantive model), SMC-FCS derives imputations that are compatible with the substantive models. SMC-FCS is implemented by `smcfcs` function in the `R` `smcfcs` package and a range of common models (e.g., linear regression, logistic regression, poisson regression, Weibull regression and Cox regression) are available.

[1] With left-tailed (MARleft), centered (MARmid), both tailed (MARtail) or right-tailed (MARright) missingness mechanism, a higher probability of $\boldsymbol{X}$ being missing are assigned to the units with low, centered, extreme and high values of Y respectively.

Results. Table 2 displays the results of the simulation, including estimates of α, β_1, β_2, σ_ϵ, R^2 and the coverage of nominal 95% confidence intervals of β_1 and β_2. In general, MPMM performed similarly to the polynomial combination method. There were no discernible biases for both approaches with five types of missingness mechanisms (MCAR, MARleft, MARmid, MARtail, and MARright). The coverage of the CIs for β_1 and β_2 from MPMM and PC was close to 95% with MCAR, MARleft, and MARmid. However, MPMM and PC had low CI coverage with MARtail and MARright. The undercoverage issue is due to the data-driven nature of predictive mean matching. PMM might result in implausible imputations when sub-regions of the sample space are sparsely observed or even truncated, possibly because of the extreme missing data mechanism and the small sample size. In such a case, two possible results may occur. First, the same donors are repeatedly selected for the missing unit in the sparsely populated sample space, which may lead to an underestimation of the variance of the considered statistic [40]. Second, more severely, the selected donors are far away from the missing unit in the sparsely populated sample space, which may lead to a biased estimate of the considered statistic.

Although *'impute, then transform'* method preserved the squared relationship, it resulted in severely biased estimates, even with MCAR. The CI coverage of β_2 was considerably poor, with all cases of missingness mechanisms. With MCAR, *'transform, then impute'* method yielded unbiased regression estimates and correct CI coverage for β_1 and β_2. However, TTI distorted the quadratic relation between $\boldsymbol{X}$ and $\boldsymbol{X^2}$. It also gave severely biased results, and the CIs for β_1 and β_2 had 0% coverage with MARleft, MARtail, and MARright. Since we knew the scientific model in the simulation study and specified a correct imputation model, SMC-FCS provided unbiased estimates and closed to 95% CI coverage with all five missingness mechanisms. Furthermore, It was noteworthy that with MARtail and MARright, MPMM and PC yielded relatively accurate estimations for σ_ϵ and R^2 compared with the model-based imputation method.

Overall, the multivariate predictive mean matching yielded unbiased estimates of regression parameters and preserved the quadratic structure between $\boldsymbol{X}$ and $\boldsymbol{X^2}$. Figure 1 shows an example of the observed data and imputed data relationships between $\boldsymbol{X}$ and $\boldsymbol{X^2}$, generated by the multivariate predictive mean matching method.

4.2 Linear Regression with Interaction Term

This section considers a linear regression substantive model, which includes two predictors and their interaction effect.

Table 2. Average parameter estimates for different imputation methods under five different missingness mechanisms over 1000 imputed datasets ($n = 5000$) with 40% missing data. The designed model is $Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$, where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$ and $\epsilon \sim N(0, 1)$. The population coefficient of determination $R^2 = .75$.

	Missingness mechanism				
	MCAR	MARleft	MARmid	MARtail	MARright
Transform, then impute					
Intercept (α)	0	0.15	−0.04	0	−0.11
Slope of X (β_1)	1(0.93)	0.93(0.02)	0.97(0.68)	1.13(0)	1.27(0)
Slope of X^2 (β_2)	1(0.92)	0.93(0)	0.96(0.13)	1.13(0)	1.27(0)
Residual SD (σ_ϵ)	1	0.96	1	1.06	1.13
R^2	0.75	0.77	0.75	0.72	0.68
Impute, then transform					
Intercept (α)	0.32	0.22	0.2	0.45	0.49
Slope of X (β_1)	0.94(0.62)	0.97(0.91)	0.89(0.08)	1(0.99)	1.04(0.92)
Slope of X^2 (β_2)	0.68(0)	0.68(0)	0.74(0)	0.62(0)	0.7(0)
Residual SD (σ_ϵ)	1.41	1.36	1.35	1.52	1.57
R^2	0.5	0.54	0.55	0.42	0.38
PC					
Intercept (α)	0	0	0	−0.05	−0.06
Slope of X (β_1)	1(0.93)	1(0.93)	1(0.93)	1(0.85)	1(0.82)
Slope of X^2 (β_2)	1.01(0.9)	1(0.94)	1(0.93)	1.07(0.12)	1.09(0.09)
Residual SD (σ_ϵ)	1	1	1	1.05	1.07
R^2	0.75	0.75	0.75	0.72	0.71
PMM-CRA					
Intercept (α)	0	0	0	−0.03	−0.03
Slope of X (β_1)	1(0.93)	1(0.93)	1(0.91)	1.04(0.47)	1.06(0.4)
Slope of X^2 (β_2)	1(0.91)	1(0.95)	1(0.93)	1.05(0.25)	1.07(0.23)
Residual SD (σ_ϵ)	1	1	1	1.05	1.07
R^2	0.75	0.75	0.75	0.72	0.71
SMC-FCS					
Intercept (α)	0.01	0	0	0.03	0.05
Slope of X (β_1)	1(0.96)	1(0.95)	1(0.95)	1(0.97)	1.01(0.97)
Slope of X^2 (β_2)	1(0.95)	1(0.96)	1(0.94)	1(0.96)	1.01(0.93)
Residual SD (σ_ϵ)	1.04	1	1	1.11	1.12
R^2	0.73	0.75	0.75	0.69	0.68

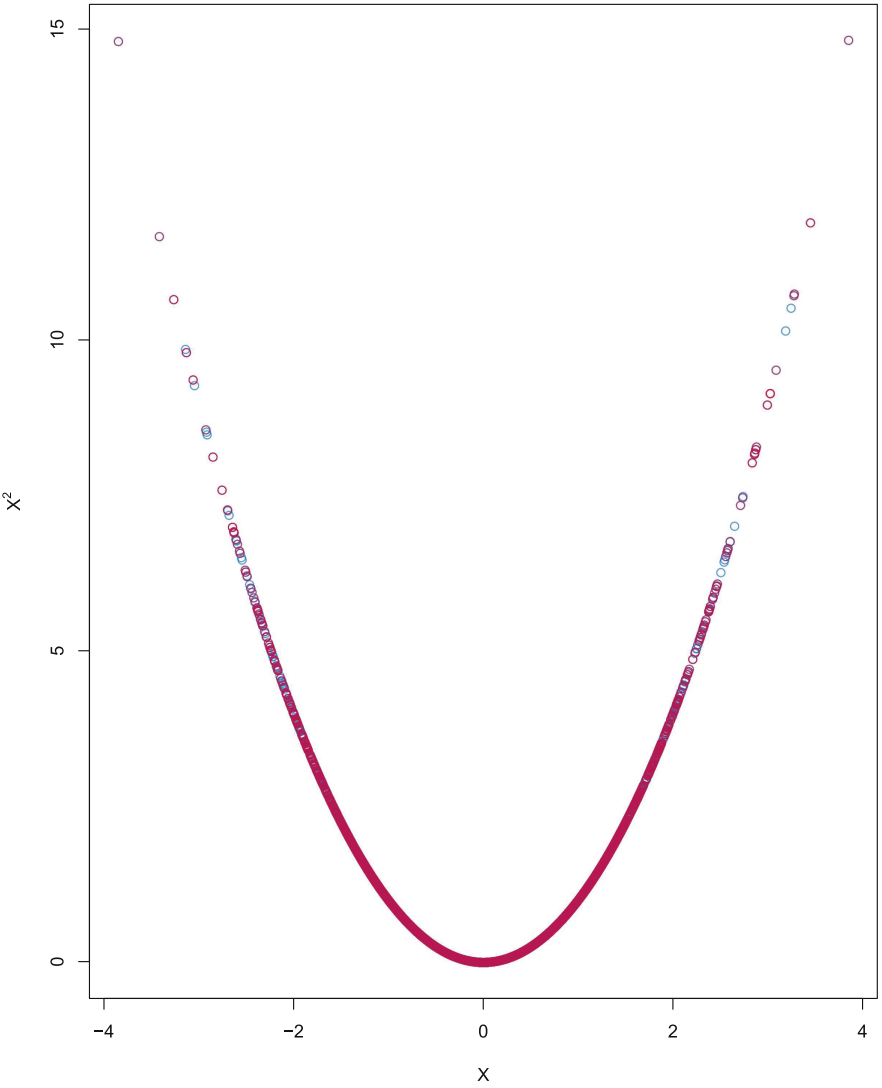

Fig. 1. Predictive mean matching based on canonical regression analysis. Observed (blue) and imputed values (red) for X and X^2.

Simulation Conditions. The dependent variable Y was generated according to the analysis model

$$\boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_1 X_2} + \epsilon \tag{5}$$

where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$, two predictors $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and error term ϵ were assumed as standard normal distributions. Under five types of missingness mechanisms: MCAR, MARleft, MARmid, MARtail, and MARright, the

probability of jointly missing X_1 and X_2 was set to 0.4. There were no units with missing values on either $\boldsymbol{X_1}$ or $\boldsymbol{X_2}$. Missing values were amputed with the `ampute` function from the package `MICE` in `R`. For each simulation scenario, $n = 5000$ units were generated and 1000 simulations were repeated.

Estimation Methods. We evaluated and compared the same methods as under Sect. 4.1, except the polynomial combination method. The model-based imputation method ensures a compatible imputation model by accommodating the designed model

Results. Table 3 shows the estimates of α, β_1, β_2, β_3, σ_ϵ, R^2 and the coverage of the 95% confidence intervals for β_1, β_2 and β_3. With MCAR, MARleft, and MARmid, MPMM was unbiased, and the CI coverage for regression weights was at the nominal level. While similar to the linear regression with quadratic term situation, with MARtail and MARright, MPMM yielded unbiased estimates but had relatively reduced confidence interval coverage. The reason is explained in Sect. 4.1.2. *'Transform, then impute'* method did not preserve the relations even though it resulted in plausible inferences in cases of MCAR and MARmid. The imputations were not plausible. Moreover, with MARleft, MARtial, and MARright, *'transform, then impute'* method gave severely biased estimates and extremely poor CI coverage. *'Impute, then transform'* method generally yielded biased estimates, and the CI for coefficients β_1, β_2 and β_3 had lower than nominal coverage with all five types of missingness. SMC-FCS yielded unbiased estimates of regression weights and had correct CI coverage in all simulation scenarios. The only potential shortcoming of the model-based imputation method was that the estimates of σ_ϵ and R^2 showed slight deviations from true values with MARtail and MARright.

4.3 Incomplete Dataset with Inequality Restriction $\boldsymbol{X_1 + X_2 \geqq C}$

Multiple predictive mean matching is flexible to model relations among missing variables other than linear regression with polynomial terms or interaction terms. Lastly, we would evaluate the inequality restriction $\boldsymbol{X_1} + \boldsymbol{X_2} \geqq C$, which is relatively difficult for the model-based imputation approach to specify. One application of such inequality restriction would be the analysis of the academic performance of qualified students. For example, the sum score of mid-term and final exams should exceed a fixed value.

Simulation Conditions. The data was generated from:

$$\begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_3} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 & 3.2 \\ 3.2 & 4 \end{pmatrix} \right],$$

$\boldsymbol{X_2} = 3 - \boldsymbol{X_1} + \epsilon$, where ϵ followed a standard uniform distribution. The sum of $\boldsymbol{X_1} + \boldsymbol{X_2} \geqq 3$ was the restriction in the generated data. We simulated

Table 3. Average parameter estimates for different imputation methods under five different missingness mechanisms over 1000 imputed datasets ($n = 5000$) with 40% missing data. The designed model is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$, where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$ and $\epsilon \sim N(0, 1)$. The population coefficient of determination $R^2 = .75$.

	Missingness mechanism				
	MCAR	MARleft	MARmid	MARtail	MARright
Transform, then impute					
Intercept (α)	0	0.05	−0.05	0.06	0.05
Slope of X_1 (β_1)	1(0.93)	0.96(0.4)	1(0.94)	1.05(0.42)	1.08(0.05)
Slope of X_2 (β_2)	1(0.94)	0.96(0.4)	1(0.96)	1.05(0.38)	1.09(0.02)
Slope of X_1X_2 (β_3)	1(0.94)	0.96(0.53)	0.95(0.25)	1.06(0.31)	1.09(0.02)
Residual SD (σ_ϵ)	1	0.97	1	1.02	1.04
R^2	0.75	0.76	0.75	0.74	0.73
Impute, then transform					
Intercept (α)	0	−0.04	−0.01	0.01	0.11
Slope of X_1 (β_1)	0.98(0.88)	1.05(0.51)	0.96(0.71)	0.98(0.9)	0.95(0.69)
Slope of X_2 (β_2)	0.98(0.88)	1.05(0.48)	0.96(0.73)	0.98(0.92)	0.95(0.69)
Slope of X_1X_2 (β_3)	0.64(0)	0.64(0)	0.7(0)	0.54(0)	0.61(0)
Residual SD (σ_ϵ)	1.25	1.18	1.22	1.28	1.37
R^2	0.61	0.65	0.63	0.59	0.53
PMM-CRA					
Intercept (α)	0	0	0	0.01	0.02
Slope of X_1 (β_1)	1(0.93)	1(0.86)	1(0.92)	1.02(0.8)	1.02(0.73)
Slope of X_2 (β_2)	1(0.93)	1(0.84)	1(0.93)	1.02(0.8)	1.02(0.77)
Slope of X_1X_2 (β_3)	1(0.94)	1.01(0.86)	1(0.93)	1.02(0.71)	1.03(0.68)
Residual SD (σ_ϵ)	1	1.01	1	1.03	1.03
R^2	0.75	0.74	0.75	0.74	0.74
SMC-FCS					
Intercept (α)	0	−0.01	0	0.01	0.03
Slope of X_1 (β_1)	1(0.95)	1.01(0.95)	1(0.95)	1(0.96)	0.99(0.95)
Slope of X_2 (β_2)	0.99(0.94)	0.99(0.93)	1(0.97)	1(0.96)	0.99(0.96)
Slope of X_1X_2 (β_3)	1(0.95)	1(0.96)	1(0.95)	1(0.97)	1.01(0.93)
Residual SD (σ_ϵ)	1.02	1.02	1	1.07	1.06
R^2	0.74	0.74	0.75	0.71	0.72

missingness such that rows in the block ($\boldsymbol{X_1, X_2,}$) were always either observed or completely missing. We considered 30% joint missingness of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. 2000 subjects were generated and 1000 simulations were performed for two missingness

mechanisms : MCAR and MARright. We evaluated the mean of X_1 and X_2 and the coverage of nominal 95% CIs.

Estimation Methods. We compared MPMM with PMM to illustrate the limited performance of univariate imputation approaches when there are relations connected to multiple missing variables. We did not apply joint modeling and univariate model-based imputation methods because it is hard to specify the designed inequality restriction.

Table 4. Average parameter estimates for MPMM and PMM under MCAR and MARright over 1000 imputed datasets ($n = 2000$) with 30% missing data. The designed model is introduced in Sect. 4.3.1. The true values of $E(X_1)$ and $E(X_2)$ are 0 and 3.5.

	MPMM				PMM			
	MCAR		MARright		MCAR		MARright	
	MEAN	COVERAGE	MEAN	COVERAGE	MEAN	COVERAGE	MEAN	COVERAGE
$E(X_1)$	0	0.95	0.01	0.94	0	0.92	−0.3	0
$E(X_2)$	3.5	0.95	3.51	0.95	3.5	0.92	3.8	0

Results. Table 4 shows the mean estimates of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ and coverage of the corresponding 95% CIs. The true values for $E(\boldsymbol{X_1})$ and $E(\boldsymbol{X_2})$ are 0 and 3.5. MPMM yielded unbiased estimates with MCAR and MARright and had the correct CI coverage. However, PMM was unbiased with close to 95% when the missingness mechanism is MCAR. It had considerable bias and extremely poor coverage with MARright. The reason is that the relations between X_1 and X_2 are not modeled [29].

5 Conclusion

Predictive mean matching is an attractive method for missing data imputation. However, because of its univariate nature, PMM may not keep relations between variables with missing units. Our proposed modification of predictive mean matching, MPMM, is a multivariate extension of PMM that imputes a block of variables. We combine canonical regression analysis with predictive mean matching so that the models for donors selection are appropriate when there are restrictions involving more than one variable. MPMM could be valuable because it inherits the advantages of predictive mean matching and preserves relations between partially observed variables. Moreover, since predictive mean matching performs well in a wide range of simulation studies, so can the multivariate predictive mean matching.

We assess the performance of the multivariate predictive mean matching under three different substantive models with restrictions. In the first two simulation studies, MPMM provides unbiased estimates where the scientific model

includes square terms and interaction terms under both MCAR and MAR missingness mechanisms. However, with MARtail and MARright, MPMM suffers the undercoverage issue because the density of the response indicator is heavy-tailed with our simulation setup. It makes units with large Y almost unobserved and more missing than observed data in the tail region. The missingness mechanism is commonly moderate in practice, unlike MARtial and MARright in simulation studies. Overall, when no sub-regions of the sample space are sparsely observed, the multiple predictive mean matching analysis will provide unbiased estimates and correct CI coverage.

SMC-FCS yields better estimates and CI coverage of regression weights, but MPMM provides relatively accurate σ_ϵ and R^2. The comparison is not entirely fair because SMC-FCS, as used here, requires the correct substantive model for the data. In practice, we often do not know the model, and MPMM becomes attractive. MPMM is an easy-to-use method when increasing variables in the datasets or only the estimates are of interest.

The third simulation shows the appealing properties of MPMM. When relations of missing variables are challenging to model, MPMM becomes the most effective approach to imputation. We expect that MPMM could be applied to other relations not yet discussed in Sect. 4.

We limited our calculations and analyses to normal distributed $\boldsymbol{X}$. However, since Vink [30] concluded that PMM yields plausible imputations with non-normal distributed predictors, we argue that distributions of predictors will not significantly impact the imputations. We focus on the simple case with one missing data pattern. One possible way to generalize MPMM to more complicated missing data patterns is proposed in the appendix. The general idea is to partition the cases into groups of identical missing data patterns in the block imputed with MPMM. We then perform the imputation in ascending order of the fraction of missing information, i.e., we first impute cases with relatively small missing data problems. Considering to impute partially observed covariates for linear regression with a quadratic term $Y = X + X^2$, we first impute cases with only missing value in X^2 by square the observed X. Then cases with only missing value in X are imputed with one square root of $Y = X + X^2$. However, the selection of roots should be modeled with logistic regression. Finally, we impute cases with jointly missing X and X^2 with MPMM. The comprehensive understanding of MPMM with multiple missing data patterns is an area for further research.

Appendix

The MPMM algorithm with multiple missing patterns:

1. Sort the rows of $\boldsymbol{Y}$ into S missing data patterns $\boldsymbol{Y}_{[s]}, s = 1, \cdots, S$.
2. Initialize $\boldsymbol{Y}_{\boldsymbol{mis}}$ by a reasonable starting value.
3. Repeat for $T = 1, \cdots, t$.
4. Repeat for $S = 1, \cdots, s$.

5. Impute missing values by steps 1–8 of PMM-CRA algorithm proposed in Sect. 2.3.
6. Repeat steps 1–5 m times and save m completed datasets.

References

1. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (2004)
2. Schafer, J.L.: Analysis of Incomplete Multivariate Data. CRC Press, Boca Raton (1997)
3. Van Buuren, S.: Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med Res. **16**(3), 219–242 (2007)
4. Goldstein, H., Carpenter, J.R., Browne, W.J.: Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. J. Roy. Stat. Soc. Ser. A. **177**(2), 553–564 (2014)
5. van Buuren, S.: Flexible Imputation of Missing Data, 2nd edn. Chapman and Hall/CRC (2018). https://doi.org/10.1201/9780429492259
6. Little, R.J.A.: Missing-data adjustments in large surveys. J Bus. Econ. Stat. **6**(3), 287–296 (1988). https://doi.org/10.1080/07350015.1988.10509663
7. van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. J. Stat. Softw. **45**(3) (2011). https://doi.org/10.18637/jss.v045.i03
8. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, vol. 793. Wiley, New York (2019)
9. Schafer, J.L.: Multiple imputation: a primer. Stat. Methods Med. Res. **8**(1), 3–15 (1999)
10. Sinharay, S., Stern, H.S., Russell, D.: The use of multiple imputation for the analysis of missing data. Psychol. Methods **6**(4), 317 (2001)
11. Allison, P.D.: Missing Data. Sage Publications, Thousand Oaks (2001)
12. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147 (2002)
13. Longford, N.: Multilevel analysis with messy data. Stat. Methods Med. Res. **10**(6), 429–444 (2001)
14. Olinsky, A., Chen, S., Harlow, L.: The comparative efficacy of imputation methods for missing data in structural equation modeling. Eur. J. Oper. Res. **151**(1), 53–79 (2003)
15. Allison, P.D.: Missing data techniques for structural equation modeling. J. Abnorm. Psychol. **112**(4), 545 (2003)
16. Twisk, J., de Vente, W.: Attrition in longitudinal studies: how to deal with missing data. J. Clin. Epidemiol. **55**(4), 329–337 (2002)
17. Demirtas, H.: Modeling incomplete longitudinal data. J. Mod. Appl. Stat. Methods **3**(2), 5 (2004)
18. Pigott, T.D.: Missing predictors in models of effect size. Eval. Health Prof. **24**(3), 277–307 (2001)
19. Schafer, J.L.: Multiple imputation in multivariate problems when the imputation and analysis models differ. Stat. Neerl. **57**(1), 19–35 (2003)
20. Seaman, S.R., White, I.R.: Review of inverse probability weighting for dealing with missing data. Stat. Methods Med. Res. **22**(3), 278–295 (2013)

21. Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R., Herring, A.H.: Missing-data methods for generalized linear models: a comparative review. J. Am. Stat. Assoc. **100**(469), 332–346 (2005)
22. Israels, A.Z.: Eigenvalue Techniques for Qualitative Data (m&t series). DSWO Press, Leiden (1987)
23. Izenman, A.J.: Reduced-rank regression for the multivariate linear model. J. Multivar. Anal. **5**(2), 248–264 (1975)
24. Sun, L., Ji, S., Yu, S., Ye, J.: On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In: Twenty-First International Joint Conference on Artificial Intelligence (2009)
25. McDonald, R.P.: A unified treatment of the weighting problem. Psychometrika. **33**(3), 351–381 (1968). https://doi.org/10.1007/bf02289330
26. Rubin, D.B.: Statistical matching using file concatenation with adjusted weights and multiple imputations. J. Bus. Econ. Stat. **4**(1), 87 (1986). https://doi.org/10.2307/1391390
27. Vink, G., Lazendic, G., van Buuren, S.: Partitioned predictive mean matching as a large data multilevel imputation technique. Psychol. Test Assess. Model. **57**(4), 577–594 (2015)
28. Heitjan, D.F., Little, R.J.A.: Multiple imputation for the fatal accident reporting system. Appl. Stat. **40**(1), 13 (1991). https://doi.org/10.2307/2347902
29. Morris, T. P., White, I. R., Royston, P.: Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med. Res. Methodol. **14**(1) (2014). https://doi.org/10.1186/1471-2288-14-75
30. Vink, G., Frank, L. E., Pannekoek, J., van Buuren, S.: Predictive mean matching imputation of semicontinuous variables. Stat. Neerl. **68**(1), 61–90 (2014). https://doi.org/10.1111/stan.12023
31. Carpenter, J., Kenward, M.: Multiple Imputation and Its Application. Wiley, New York (2012)
32. Seaman, S.R., Bartlett, J.W., White, I.R.: Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med. Res. Methodol. **12**(1) (2012). https://doi.org/10.1186/1471-2288-12-46
33. Rencher, A.C.: Methods of Multivariate Analysis, vol. 492. Wiley, New York (2003)
34. Van Den Wollenberg, A.L.: Redundancy analysis an alternative for canonical correlation analysis. Psychometrika **42**(2), 207–219 (1977)
35. Schouten, R.M., Lugtig, P., Vink, G.: Generating missing values for simulation purposes: a multivariate amputation procedure. J. Stat. Comput. Simul. **88**(15), 2909–2930 (2018). https://doi.org/10.1080/00949655.2018.1491577
36. R Core Team: R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria (2021). https://www.R-project.org/
37. Von Hippel, P.: How to impute interactions, squares, and other transformed variables. Sociol. Methodol. **39**(1), 265–291 (2009)
38. Vink, G., van Buuren, S.: Multiple imputation of squared terms. Sociol. Methods Res. **42**(4), 598–607 (2013). https://doi.org/10.1177/0049124113502943
39. Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., Initiative*, A.D.N.: Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat. Methods Med. Res. **24**(4), 462–487 (2015)
40. de Jong, R., van Buuren, S., Spiess, M.: Multiple imputation of predictor variables using generalized additive models. Commun. Stat. Simul. Comput. **45**(3), 968–985 (2014). https://doi.org/10.1080/03610918.2014.911894