# Dissertation

Mingyang Cai

# Acknowledgements

I wish to express my gratitude to the people who have been instrumental in completing this project. I would like to show my greatest appreciation to my supervisors. Stef, I thank you for your valuable guidance and advice. Without your vast knowledge of statistics, this project would not have been possible. You inspired me greatly to research multiple imputations and has been critical to my scientific growth. Deep gratitude is also due to Gerko, my daily supervisor. Two meetings every week, parties held at your home and tips for living in the Netherlands, all these things make me feel at home away from home. I learn a lot from your elegant programming. I still remember your words at our first meeting: the PhD project is to prove your ability to conduct research independently. This word really leads the way to my project. I was pleased to have supervision from both of you.

I thank everyone once joined in the weekly mice meeting. The shared works and ideas, the discussed questions during the meeting really have an influence on me. Furthermore, I am grateful to everyone from M&S at Utrecht University. I enjoyed various kinds of research meetings and activities here.

My friends and family are important to me. I thank my parents for giving me the opportunity to learn more about statistics and develop skills to conduct scientific research. You could not have given me a greater gift. I also thank my brother and friends to make a joyful life.

# Contents

# Chapter 1

# Introduction

Missing data commonly occur in scientific research and have a significant impact on the downstream analysis. First, most statistical analyses require complete data and can not apply to incomplete data directly. Second, the reduced sample size arose from missing data generally induces estimates with less efficiency and results with a loss of statistical power [Rubin, 1987].

For instance, in clinical research, investigators may encounter difficulties of prognostic covariates measurement, such as estrogen receptor status, nodal status, or size of the tumour in cancer studies; CD4 counts in AIDS studies [Ibrahim et al., 2012]. Such missing variables restrict the assessment of balance in covariate distributions in observational studies. In social science, longitudinal studies are used to investigate some developmental trends over an extended period. However, longitudinal studies suffer attrition of participants. Missing data occur when participants drop out before the end of the study. In sample survey research, respondents may be assigned to a subset of all questions by design. Additionally, they may refuse to answer specific questions which involve private information (e.g., personal income) or cost much time (e.g., watching one movie before answering questions).

Although missing data annoy scientists and researchers, the field of missing data analysis has been dramatically developed. A large number of approaches were proposed to handle missing data with different features. The general idea of missing data methods is to deduce information about missing data from what we observed.

## 1.1 Missingness mechanisms

It is necessary to understand the reason for non-response. Rubin [1976] assumed that the mechanisms of missing data could be modelled and defined three categories of missing data problems based on different missingness mechanisms. If the probability of variables being missing is equal for all cases, then the variables are missing completely at random (MCAR). An example of MCAR is when a random sample is drawn from the population, each individual has the same probability of being sampled. Individuals who are not included in the sample could be viewed as MCAR. In such a case, the observed data is still representative of the population, which implies the validity of complete data analysis. Although simplifying missing data problems, MCAR is often impractical. If the probability of variables being missing depends on the observed data, then the variables are missing at random (MAR), which is more general than MCAR. An example of MAR is that younger individuals have a higher probability of being sampled from the population, where the age of the collected sample is completely observed. Most missing data analyses are developed based on the MAR assumption because the MAR implies an ignorable missingness mechanism [Little and Rubin, 2019]. In such a case, the unobserved data share identical distributions as the observed data, which means that the model generated by the observed data could be applied to the missing data. However, it is notable that ignorable missingness mechanisms do not entirely disregard the missingness mechanisms. The imputer should consider variables that explain the probability of being missing. Sometimes, the missingness mechanisms are scientific interests. Suppose the probability of variables being missing depends on partially observed data or unobserved data. Then, the variables are missing not at random (MNAR), which is a more complex situation to handle the missing data problem. MNAR means that the research has no information about the missingness mechanism. An example of MNAR is that a researcher would like to analyse the income of a population. However, people with higher income may tend to refuse income disclosure.

The classification of missingness mechanisms inspires researchers to understand the reason for missing data occurrence and provide information about which missing data analyses will produce valid inferences. For instance, in general, complete case analysis only works under MCAR.

## 1.2 Missing data methods

There are mainly four categories of missing data approaches proposed in the literature: complete-case analysis, weighting procedures, direct maximum likelihood methods and multiple imputation methods [Little and Rubin, 2019]. The four methods are not mutually exclusive. This section will give an overview of the first three methods. Since this dissertation focuses on multiple imputation (MI), a more detailed introduction of MI will be given later.

Complete case (CC) analysis, also named list-wise deletion, discards cases with missing values and only analyses complete observations. Although CC analysis is a straightforward and convenient solution to the missing data problem, it yields unbiased but less efficient mean vectors, covariance matrix and regression weights merely under MCAR. Sometimes, it may also produce acceptable results when there are a minor amount of partially observed cases. However, CC analysis wastes a large amount of data, which leads to loss of precision and bias when missingness mechanisms are MAR or MNAR.

Weighting methods commonly apply for unit non-response, which means that all outcomes data for the respondent who refuses to participate in the survey are missing. The general idea of weighting analysis is analogous to weighting in randomisation inference. The weighting methods define the probability of selection ($\phi$) in the sample and apply the classical Horvitz - Thompson estimator [Horvitz and Thompson, 1952] to evaluate scientific interest. Suppose $y_i$ is the value of the variable $Y$ for unit $i$ ($i = 1, 2, \dots, n$), the Horvitz - Thompson estimator of the population mean is $\bar{Y}_{HT} = \sum_{i=1}^{n} \phi^{-1} y_i / n$.

When there are non-responses units, the probability of response $\hat{p}$ should be considered and the adjusted estimator is:

$$\frac{\sum_{i=1}^{n}(\phi\hat{p}_i)^{-1}y_i}{\sum_{i=1}^{n}(\phi\hat{p}_i)^{-1}}$$

Weighting methods are a conceptually and computationally convenient procedure to reduce bias from CC analysis. However, although it takes account of missingness mechanisms, the application of weighing methods is limited because of no full use of incomplete observed units and less straightforward computations of variances.

Direct maximum likelihood methods include extensive procedures that define a model for the complete data, estimate the parameters of the model with maximum likelihood and conduct statistical inferences. Little and Rubin [2019] developed sophisticated direct maximum likelihood methods to estimate parameters of likelihood, obtain standard errors from information matrix and create missing data. For example, suppose a random variable $Y$ consists of observed and missing parts ($Y_{obs}$ and $Y_{mis}$). Then, with the ignorable missingness mechanism assumption, the likelihood for the parameters of the imputation model ($\theta$) is an integral over the missing values:

$$L(\theta|Y_{obs}) = \int f_Y(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}.$$

Direct maximum likelihood methods could provides unbiased and efficient estimates under both MCAR and MAR.

## 1.3 Multiple imputation

Multiple imputation (MI) is a general approach for the analysis of incomplete datasets. It involves generating several plausible imputed datasets and aggregating different results into a single inference. First, missing cells are filled with synthetic data drawn from corresponding posterior predictive distributions. Next, this procedure is repeated multiple times, resulting in several imputed datasets. The scientific interests are then estimated for each imputed dataset by complete-data analyses. Finally, different estimates are pooled into one inference using Rubin's rule, which accounts for within and across imputation uncertainty [Rubin, 1987].

Rubin proposed the main principles of MI at the end of the 70's, but the uptake and development of MI techniques was growing at a favorable rate over the last two decades. Nowadays, various technologies of MI are generated for different statistical models, for instance, multilevel multiple imputation [Longford, 2001], MI for longitudinal data [Twisk and de Vente, 2002, Demirtas, 2004], MI for structural equation modeling [Olinsky et al., 2003, Allison, 2003]. Because of the methodological development of MI, the techniques are applied to a wide range of fields (e.g., epidemiology [Mueller et al., 2008], politics [Tanasoiu and Colonescu, 2008], genetics [Souverein et al., 2006], psychology [Sundell et al., 2008] and sociology [Finke and Adamczyk, 2008]) and implemented in many software packages (e.g. `mice` and `mi` in `R`, `IVEWARE` in `SAS`, `ice` in `STATA` and module `MVA` in `SPSS`) [van Buuren and Groothuis-Oudshoorn, 2011, Azur et al., 2011].

There are two general approaches for imputing multivariate data: joint modelling (JM) and fully conditional specification (FCS). Joint modeling imputation assumes a model $p(Y^{mis}, Y^{obs} \mid \theta)$ for the complete data and a prior distribution $p(\theta)$ for the parameter $\theta$. Joint modelling partitions the observed data into groups based on the missing pattern and imputes the missing data within each missing pattern according to corresponding predictive distribution. Under the assumption of ignorability, the parameters of the predictive distribution for different missing patterns are generated from the posterior joint distribution. Schafer [1997] proposed joint modelling methods for multivariate normal data, categorical data and mixed normal-categorical data. Joint modelling approaches have solid theoretical properties (i.e., compatibility between the imputation and substantive models), while it lacks the flexibility of model specification.

Fully conditional specification specifies the distribution for each partially observed variable conditional on all other variables $P(Y(j)|Y(-j), X, \theta_j)$ and imputes each missing variable iteratively. The FCS starts with naive imputations such as a random draw from the observed values. The $t$th iteration for the incomplete variable $Y(j)$ consists of the following draws:

$$\begin{aligned}\theta_j^t &\sim f(\theta_j) f(Y^{obs}(j) | Y^{t-1}(-j), X, \theta_j) \\ Y^{mis(t)}(j) &\sim f(Y^{mis}(j) | Y^{t}(-j), X, \theta_j^t),\end{aligned}$$

where $f(\theta_j)$ is generally specified as a noninformative prior. After a sufficient number of iteration, typically with 5 to 10 iterations [van Buuren, 2018, Oberman et al., 2020], the stationary distribution is achieved. The final iteration generates a single imputed dataset and the multiple imputations are created by applying FCS in parallel $m$ times. Since FCS provides tremendous flexibility in specifying imputation models for multivariate partially observed data, FCS is now a widely accepted and popular MI approach [Van Buuren, 2007]. Even while, FCS lacks a satisfactory theory and has a potential risk in incompatibility.

Hybrid imputation, also called block imputation, combines the flexibility of FCS with the attractive theoretical properties of JM. A block consists of one or more variables. If the block has multiple variables, then multivariate imputation methods will be applied to impute those variables jointly. The joint modelling approach is the case where all variables form one block, while the FCS approach treats each variable as a separate block. When the imputation model of one variable is potentially incompatible, or when its theoretical properties are not thoroughly studied, hybrid imputation would merge that variable with other variables and apply the joint modelling imputation approach to that block.

On the other hand, when the joint distribution of several missing variables is ambiguous, hybrid imputation could use the FCS approach to impute each variable. In general, the apparent advantage of hybrid imputation is the flexibility of model specification. However, hybrid imputation methods are hardly known or studied.

## 1.4 Aim

This dissertation develops hybrid imputation in both methodological and practical aspects. Based on missingness patterns and restrictions on the data, different block-wise partition strategies and the corresponding block imputation methods are considered to provide plausible imputations. Although hybrid imputation is a broad and undeveloped field, it could be a more user-friendly and flexible strategy for multivariate imputation. This dissertation addresses the following hybrid imputation problems.

First, in medical and epidemiological research, the analysis model commonly contains squared terms [Seaman et al., 2012, Bartlett et al., 2015]. In order to generate plausible imputations, the MI procedure should preserve the relationship between the original variable and its squared counterpart. Vink and van Buuren [2013] proposed a hot-deck multiple imputation method, named polynomial combination (PC) method, for imputation models containing squared terms. The method yields unbiased regression estimates and preserves the quadratic relationships in the imputed data for both MCAR and MAR mechanisms. However, the coverage rate of the PC method is not thoroughly studied. Chapter 2 dives into the coverage rate of the PC method and proposes a minor adjustment to the PC method. As a result, the performance of the PC method is improved to some extent.

Second, in many domains of statistics, restrictions among variables are not limited to the quadratic effect, for instance, data transformations, interactions and range restrictions. Therefore, it is sensible to embrace all relations of scientific interest in the imputation model. With hybrid imputation strategies, the researcher could set variables involved in a restriction into one block and perform a joint imputation. Chapter 3 proposes a multivariate predictive mean matching (MPMM), which generalises univariate predictive mean matching [Rubin, 1986, Little, 1988] to impute multiple variables simultaneously.

Third, causal inference is widely used in epidemiology, biology and social science. The treatment effect is usually calculated by the average treatment effect. However, to provide more accurate treatment recommendations, the analyst should take account of the heterogeneity of treatment effects across individuals, also known as the individual treatment effect. Because only one of the potential outcomes is observed for each individual, the causal inference could also be viewed as a missing data problem. Chapter 4 proposes a hybrid imputation method that sets potential outcomes in one block and imputes them to estimate individual treatment effects.

Fourth, the missing data and parameters of imputation models are often generated from the corresponding posterior distribution. However, the prior is limited to the non-informative setting. Multiple imputation with informative prior is not well developed. Open problems could be: How to implement MI with informative prior elegantly; The compatibility of FCS with informative prior; What is the corresponding priors of the substantive joint distribution when the compatibility of FCS with informative prior holds; How could MI with informative prior be applied in practice. Chapter 5 investigates the compatibility of univariate imputation under normal linear regression with normal inverse-gamma priors.

Finally, a critical part of the multiple imputation process is selecting sensible models to generate plausible values for the missing data. The validity of complete data analysis on imputed datasets relies on the congeniality of the imputation model and the substantive model of interest [Meng, 1994]. A comprehensive overview of model diagnostic in multiple imputation is available in Nguyen et al. [2017]. Chapter 6 proposes a novel strategy to diagnose multiple imputation models based on posterior predictive checking.

## 1.5 Workflow of `MICE`

A straightforward implementation of hybrid imputation can be found in the `MICE` algorithm proposed by van Buuren and Groothuis-Oudshoorn [2011]. Version 3.0 of `MICE` added a new block argument with which the user could partition the variables into blocks. The algorithm of `MICE` will iterate over blocks rather than variables. For the block that contains one variable, all imputation methods developed under the fully conditional specification framework are available. For the block that contains multiple variables, the `MICE` algorithm allows calling joint modelling imputation methods from other packages. For example, **mice::jomoImpute** is a wrapper around the **jomoImpute** function from the `mitml` package. This function is used for joint modelling imputation of multilevel data.

The MICE package creates functions for three components: imputation, analysis, and pooling. Imputed datasets are generated with function **mice()**. Complete data analysis are performed on every imputed dataset by **with()** function and combined into a single inference with function **pool()**. The software stores the output of each step in a particular class: **mids**, **mira** and **mipo**.

The MICE algorithm starts with filling missing cells by randomly drawing values from the observed data. Subsequently, incomplete variables are imputed in a block-by-block iteration — a single iteration of the algorithm cycles through all specified blocks.

The number of iterations (argument **maxit** in **mice** function) and imputations (argument **m** in **mice** function) are of importance in MI. In general, a low number of iterations appear to be enough [Brand, 1999, Van Buuren et al., 1999]. However, slow convergence may occur if there is a large amount of missing data or high autocorrelation between the imputation iterations. Oberman et al. [2020] performed a simulation study and concluded that five to ten iterations are enough for inferential validity. The convergence of the MICE algorithm could be investigated by plotting statistics of interest in each imputation chain. If no definite trends appear, convergence is achieved.

The default amount of imputations in MICE is set to be five. Although it may be beneficial to produce a higher number of imputed datasets [Royston, 2004, Graham et al., 2007, Bodner, 2008, White et al., 2011], Schafer and Olsen [1998] showed that in many cases, there is no remarkable advantage to pooling more imputed datasets. An immense amount of imputations implies more data storage and computational intensity. However, it may not be a substantive issue with the development of computer hardware.

## 1.6 Model-based imputation

The aim of model-based imputation is to find predictive models for incomplete variables. With the hybrid imputation strategy, both the univariate imputation model for one incomplete variable and the multivariate imputation model for a set of incomplete variables could be fitted based on how incomplete variables are partitioned. The model-based approaches have a solid underpinning for theory. Joint modelling procedures are commonly implemented by model-based imputation, for example, multivariate normal distribution for a set of incomplete continuous variables, multinomial distribution for a set of incomplete categorical variables and general location model for a set of incomplete mixed variables [Schafer, 1997].

For fully conditional specification, common model-based imputation procedures are available for substantive models based on regression models for continuous and discrete outcomes and the proportional hazards model. If necessary, the research could fit more complex parametric imputation models, such as hierarchical models and time series models. Since compatibility is a potential weakness of FCS, it is feasible to assess whether the iterative distribution of a set of univariate imputation models converges to a joint distribution with explicit imputation models. If not, the imputations may differ according to different orders in which missing variables are updated. The phenomenon is called the order effect.

There are two strategies to set up the imputation models. The first one, explained by Meng [1994], is that assuming an imputation model, the researcher should assess whether the analysis model is congenial to the imputation model. The second one, proposed by Bartlett et al. [2015], is that the researcher should start with the analysis model and then find an imputation model which is congenial to the analysis model. Both methods highlight the importance of the match between the substantive model and the imputation model. If there are solid scientific models for the incomplete data, model-based imputation will ensure the generated values are compatible with the substantive models. Otherwise, if a broad range of candidate models fit the data, or if there is no convinced substantive model, it will be challenging to find an imputation model to accommodate every substantive model.

For example, to estimate individual causal effects, we assume a multivariate normal distribution for the continuous potential outcomes, the validity of which is illustrated by Imbens and Rubin [2015]. In such a case, we generally partition potential outcomes and covariates into separate blocks. The continuous potential outcomes are drawn from a conditional joint normal distribution while the incomplete covariates could be imputed with fully conditional models.

## 1.7 Data-based imputation

The most widely used data-based imputation method is predictive mean matching (PMM), proposed by Rubin [1986] and Little [1988]. Although there are other non-parametric imputation methods such as tree-based imputation methods (e.g., classification and regression trees and random forest), we mainly develop imputation methods based on PMM and overview it.

PMM calculates the estimated value for the observed and unobserved part in an incomplete variable through the Bayesian normal linear regression. First, the procedure selects a set of candidate donors from all complete cases by minimizing the distance between the predicted value of the missing unit and the predicted value of all observed units. Then, the unobserved value is imputed by randomly drawing one of the observed values of the candidate donors [van Buuren, 2018].

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 [van Buuren and Groothuis-Oudshoorn, 2011, Heitjan and Little, 1991, Morris et al., 2014, Vink et al., 2014,0]. PMM is applicable to missing variables at all measurement levels. The imputation produced by PMM would always fall in the range of observed values, follow the distributions of observed data and preserve the restriction on the data. More specifically, if 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. When applying PMM, the researcher could avoid data transformations to accommodate the assumption of the parametric imputation model. For example, imputed with Bayesian normal linear regression, the missing outcome should be transformed to validate the homoscedastic and normal assumptions. Furthermore, there is no need to post-processing the imputed values to satisfy intrinsic restrictions in the data. However, since PMM is a univariate imputation method, it cannot ensure the restrictions among a set of variables.

When there is no definite substantive model, data-based imputation is an excellent alternative to produce plausible imputations. However,non-parametric imputation methods are not a substitute for sloppy modelling. When the procedure of drawing imputed values from the observed values is not reasonable, the derived imputed dataset can yield poor inference. Another potential issue for data-based imputation methods is that they cannot extrapolate beyond the range of the data and may create implausible imputations if the data is sparsely observed in some space.

In short, PMM and other kinds of data-based imputation methods are a proper choice if the researchers think the fitted parametric model is flawed, and the validity of inference will weaken drastically. The methods work better with large samples and provide imputations that capture many patterns of the complete data.

## 1.8 Imputation models evaluation

Assessing the fitness of the imputation model is also important in the MI procedure. Poor specification of the imputation model may lead to inconsistent imputed values and invalid estimates of target statistics. Nguyen et al. [2017] introduced currently available approaches for imputation models diagnostics such as comparison between the imputed value and observed data, cross-validation techniques and posterior predictive checking on scientific interests. The `MICE` package contains several graphical functions, such as **bwplot()**, **stripplot()**, **densityplot()** and **xyplot()**, to produce stripplot, box and whiskers plot, density plot and point plot of observed and imputed data to identify whether distributional discrepancy between observed and imputed data appears.

## 1.9 Outline of the dissertation

This dissertation investigates hybrid imputation in both methodological and practical aspects and is divided into three parts. Part one considers data-based imputation strategies to jointly impute a block of variables so that data restrictions can be preserved. More specifically, two methods discussed in part one are generalizations of PMM. Part two considers model-based imputation methods, which assume a multivariate normal distribution for a block of incomplete variables. The compatibility of the joint model and the corresponding univariate conditional distributions for informative priors is evaluated. Part three focuses on imputation model checking instead of the imputation procedure because it is necessary to assess the validity of selected imputation models. The discussed model diagnostic approach works for both model-based and data-based imputation methods.

# Part I

# Non-parametric imputation methods

# Chapter 2

# A note on imputing squares via polynomial combination approach

## Abstract

The polynomial combination (PC) method, proposed by Vink and Van Buuren, is a hot-deck multiple imputation method for imputation models containing squared terms. The method yields unbiased regression estimates and preserves the quadratic relationships in the imputed data for both MCAR and MAR mechanisms. However, Vink and Van Buuren never studied the coverage rate of the PC method. This paper investigates the coverage of the nominal 95% confidence intervals for the polynomial combination method and improves the algorithm to avoid the perfect prediction issue. We also compare the original and the improved PC method to the substantive model compatible fully conditional specification method (SMC-FCS) proposed by Bartlett et al. and elucidate the two imputation methods' characters.

## 2.1 Introduction

Squared terms are often included in real-life data models to accommodate some form of nonlinearity. When the analysis model contains the partially observed covariates and corresponding squared terms, some challenges arise:

1. The analysis and imputation models should accommodate squared terms, i.e., the squares themselves should be considered in the imputation procedure of corresponding linear terms.

2. The relation between the square term and its lower-order polynomial should be preserved.

3. The analysis model parameter estimates should be unbiased.

To obtain unbiased estimates, one could impute the squared term as if it were another variable. We will refer to this as *Transform, then Impute* (TTI) [Von Hippel, 2009]. However, this approach distorts the relationship between the original variable and its square. A straightforward process to preserve the quadratic relation during imputation is calculating the squared term only after imputation (*Impute, then Transform*, ITT). However, ITT biases estimates of regression coefficients, as its contribution during imputation is ignored [Von Hippel, 2009, Vink and van Buuren, 2013]. Moreover, both these partial fixes only work when the missingness is completely random [Seaman et al., 2012].

To solve these issues for a more general class of missingness mechanisms, Vink and van Buuren [2013] propose to impute the combination of the original variable and its square and decompose it into distinct roots. This polynomial combination approach (PC) is built around predictive mean matching, a nonparametric imputation hot-deck technique that does not assume a specific distribution for the data [Rubin, 1986, Little, 1988]. Seaman et al. [2012] demonstrated that predictive mean matching gives biased estimation when the analysis is a linear regression with a quadratic term and the missingness mechanism is missing at random. However, PC yields unbiased estimates for MCAR and MAR missingness mechanisms by applying a reasonable donor selection procedure on the polynomial combination instead.

More recently, Bartlett et al. [2015] proposed a substantive model compatible approach (SMC-FCS), which generalizes the imputation of nonlinear covariates beyond the squared term model. The SMC-FCS technique is efficient but needs the correct data analysis model during imputation to obtain draws of values that conform to this model. It yields unbiased estimates if 1) the substantive model is correctly specified and 2) the normality assumption of missing variables with quadratic effects is tenable. When the missing variable with quadratic effects does not follow the normal distribution, one could apply an appropriate transformation to make the normality assumption plausible (e.g., log-normal distribution). More interestingly, Bartlett et al. [2015] suggest that the SMC-FCS estimates can yield unbiased inference, meaning that estimates are both unbiased and properly covered cf. Neyman [1934]. Such an investigation into coverage of multiply imputed parameters was not part of the study by Vink and van Buuren [2013].

We now have two techniques that seem promising in imputing squared terms: the polynomial combination method and SMC-FCS. Both approaches have appealing properties, preserve the relationship between the square and its base, and yield unbiased estimates. However, both techniques differ fundamentally in their approach. SMC-FCS is a strictly model-based technique that requires the correct specification of the complete-data model and the substantive model. On the other hand, PC is a hot-deck technique with a data-driven estimation procedure that only requires the specification of the polynomial combination. We highlighted the most used properties and promising methods in Table 2.1 [Von Hippel, 2009, Vink and van Buuren, 2013, Bartlett et al., 2015].

| | `TTI` | `ITT` | `PC` | `SMC-FCS` |
|---|---|---|---|---|
| Unbiased estimates of $\beta$ | MCAR only | - | MCAR & MAR | MCAR & MAR |
| Quadratic relationship | Not preserved | Preserved | Preserved | Preserved |
| Coverage rate of $\beta$ | Poor | Poor | Unknown | Correct |
| Violation of normality | | | Robust | Somewhat robust |
| Model specification | | | Non-parametric | Parametric |

Table 2.1: Summary of properties of four squared term imputation methods.

The interpretation of Table 2.1, taking the SMC-FCS approach as an example, should be as follows:

1. SMC-FCS yields unbiased regression estimates $\beta$ provided that the missing mech-

anism is $MAR$;

2. SMC-FCS does *preserve* the quadratic relationship between the original variable and its square;

3. SMC-FCS produces *correct* coverage rates of corresponding regression estimates $\beta$;

4. SMC-FCS is *somewhat robust* against the violation of normality assumption of the covariates with quadratic effects;

5. SMC-FCS is a *parametric* imputation approach, which means SMC-FCS requires an explicit specified imputation model.

The “ - ” sign in a cell indicates that the method cannot produce unbiased estimates. Whether the PC method has a correct coverage rate is not thoroughly studied and will be investigated in the following section. There are four blank cells left because the TTI and ITT methods cannot produce unbiased regression estimates or preserve the quadratic relationships; it is redundant to investigate the violation of normality and model specification for them.

In this manuscript, we evaluate the performance of imputing squared terms with SMC-FCS and the PC method to investigate these techniques’ strengths and limitations in different scenarios. In the next section, we briefly discuss the SMC-FCS and PC methodology and propose a minor adjustment to the PC method.

## 2.2 Polynomial combination

In this paragraph we detail both the original polynomial combination (OPC) approach proposed by Vink and van Buuren [2013], as well as a modification that is robust against perfect prediction issues. We refer to the modified polynomial combination approach as MPC.

### 2.2.1 Original polynomial combination

Suppose the model of scientific interest is

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon \tag{2.1}$$

with $\epsilon \sim N(0, \sigma^2)$. We assume that $Y$ is complete and that $X = (X_{obs}, X_{mis})$ is partially missing.

The original polynomial combination method first performs predictive mean matching (PMM) [Little, 1988] on the combined variable $Z = X\beta_1 + X^2\beta_2$, and then decomposes $Z$ into components $X$ and $X^2$. Under the model in equation 2.1, two roots of variable $X$ are:

$$\begin{aligned} X_- &= -\tfrac{1}{2\beta_2}(\sqrt{4\beta_2 Z + \beta_1^2} + \beta_1) \\ X_+ &= \tfrac{1}{2\beta_2}(\sqrt{4\beta_2 Z + \beta_1^2} - \beta_1) \end{aligned} \tag{2.2}$$

where the discriminant $4\beta_2 Z + \beta_1^2$ should be larger than 0. For any imputed $Z$, we select either $X = X_-$ or $X = X_+$ and square it to derive the square term $X^2$.

The choice between the roots $X_-$ and $X_+$is made by random sampling, conditional on $Y$, $Z$, and their interaction $YZ$. The binary random variable $V$ is defined as 0 if $X < X_{min}$ and 1 if $X > X_{min}$, where the minimum of the parabola $X_{min} = -\beta_1/2\beta_2$. We model the probability $P(V = 1)$ by logistic regression as

$$\text{logit}P(V = 1) = Y\beta_{\text{Y}} + Z\beta_{\text{Z}} + YZ\beta_{\text{YZ}} \tag{2.3}$$

on the observed data. Under the assumption of ignorability, we apply the same model to calculate the predicted probability $P(V = 1)$ for $X_{mis}$. Finally, a random draw from the binomial distribution is made ($P = 0$ or 1), and the corresponding (negative or positive) root is selected as the imputation for each missing value $X_{mis}$.

### 2.2.2 Modification of Polynomial combination

Since we estimate binary variables $V$ in the OPC imputation procedure, it is necessary to avoid bias due to perfect prediction. When imputers apply the original polynomial combination method, perfect prediction occurs when all the observed binary variables $V_{obs}$ are equal to one (or zero). In this case, the likelihood tends to a limit as one or some regression coefficients tend to infinity, which leads to seriously implausible imputations of the binary variable $V$ [White et al., 2010].

Suppose all observed $X$ are located on the parabolic function's right arm, the perfect prediction arises. If no corrections are performed, the coefficients of the logistic function $\text{logit}P(V = 1) = Y\beta_Y + Z\beta_Z + YZ\beta_{YZ}$ will have extremely wide and flat posterior distributions, which tends to derive extremely positive or negative estimates of coefficients. Provided all observed $X$ are located on the right arm of the parabolic function, some missing values of $X$ would be addressed incorrectly on the left arm, as shown in Figure 2.1(a).

A computationally convenient approach to perfect prediction is data augmentation [van Buuren, 2018, Section 3.6.2]. We augment the data with a few extra observations and add a small weight to these observations. Specifically, suppose the data has $p$ predictors and a $k$ levels outcome, for each predictor $X_j, j = 1, \ldots, p$, we add two observations where $X_j$ equals either $\bar{x}_j - s_j$ or $\bar{x}_j + s_j$ for all levels of the response ($\bar{x}_j$ is the mean of $X_j$ and $s_j$ is the standard deviation of $X_j$). Other variables are fixed to their mean. Thus we add *2k* observations for each predictor and in total *2pk* observations to the data. A small weight $w =$ *(p + 1)/2pk* is assigned to each additional observation to restrict their impact on the imputation model identification.

To improve the polynomial combination method, we impute $V_{mis}$ by logistic regression of V given Y, Z, and YZ with the augmented data instead of the observed data. More specifically, based on the observed $V$, $Y$ and $Z$, the augmented data adds eight subjects shown in Table 2.2, with the weight 3/8, to the observed data. When the population estimation of the probability $P(V = 1)$ equals one (or zero), we expect the modified polynomial combination method would provide more plausible imputations, as shown in

|  | V | Y | Z |
|---|---|---|---|
| 1 | 1 | E($Y_{obs}$) + SD($Y_{obs}$) | E($Z_{obs}$) |
| 2 | 1 | E($Y_{obs}$) - SD($Y_{obs}$) | E($Z_{obs}$) |
| 3 | 0 | E($Y_{obs}$) + SD($Y_{obs}$) | E($Z_{obs}$) |
| 4 | 0 | E($Y_{obs}$) - SD($Y_{obs}$) | E($Z_{obs}$) |
| 5 | 1 | E($Y_{obs}$) | E($Z_{obs}$) + SD($Z_{obs}$) |
| 6 | 1 | E($Y_{obs}$) | E($Z_{obs}$) - SD($Z_{obs}$) |
| 7 | 0 | E($Y_{obs}$) | E($Z_{obs}$) + SD($Z_{obs}$) |
| 8 | 0 | E($Y_{obs}$) | E($Z_{obs}$) - SD($Z_{obs}$) |

Table 2.2: Augmented data

Figure 2.1(b).

### 2.2.3 SMC-FCS

The substantive model compatible fully conditional specification (SMC-FCS) is a parametric imputation method proposed by Bartlett et al. [2015]. In general, the missing predictor is imputed based on other predictors. A rejection sampling (e.g., Metropolis-Hastings algorithm) is used where the acceptance ratio is generated based on the likelihood of the substantive model. Suppose $\phi$ is a vector containing the coefficients of the model $f(Y|X)$ and $\theta_i, j = 1, \ldots, p$ is a vector containing the coefficients of the model $f(X_i|X_{-i})$, where $X_{-i}$ are all the other covariates excluding $X_i$. The parametric density function of the partially observed variable $X_i$ is proportional to $f(Y|X,\phi)f(X_i|X_{-i},\theta_i)$, rooted in the Bayesian rule:

$$\begin{aligned} f(X_i|X_{-i},Y) &= \tfrac{f(X_i,X_{-i},Y)}{f(Y,X_{-i})} \\ &\propto f(Y|X_i,X_{-i})f(X_i|X_{-i}). \end{aligned} \tag{2.4}$$

Since the density generally does not follow a standard parametric family, the rejection sampling is necessary to draw coefficients from the posterior distributions of $\phi$ and $\theta_i$. With the assumption of independent priors $f(\phi)$ and $f(\theta_i)$, the posterior distributions of $\phi$ and $\theta_i$ would be:

$$\begin{aligned} \phi &\sim f(Y|X_i,X_{-i},\phi)f(\phi) \\ \theta_i &\sim f(X_i|X_{-i},\theta_i)f(\theta_i). \end{aligned} \tag{2.5}$$

The statistical property of this approach is that if the substantive model $f(Y|X)$ is correctly specified, the imputation model will be congenial to the analysis model [Meng, 1994]. The lack of congeniality can sometimes produce implausible imputations that result in biased inferences in the downstream analysis [Robins and Wang, 2000].

## 2.3 Evaluation

We evaluated the average biases, the coverage of nominal 95% confidence intervals, and the average width of corresponding confidence intervals of the regression weights $\beta_1$ and $\beta_2$.

### 2.3.1 Simulation setup

The outcome $Y$ was simulated according to the scientific model:

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon \tag{2.6}$$

with $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$. The value of $\sigma_\epsilon$ varied according to different distributions of $X$ so that the coefficient of determination $R^2$ was always equal to 0.75.

The predictor $X$ was generated from a normal, a skew-normal, or a normal mixture distribution. The mean of $X$ was either 0 or 1, and the variance was 1 for all three distributions. The abscissa at the parabolic minimum was $X = -1/2$. Hence, when the location of $X$ was 0, there was a strong U-shaped association between $Y$ and $X$. If $X$ had location 2, the relationship between $Y$ and $X$ would be somewhat linear. For the skew-normal distribution, we set the slant parameter to be 6 when the mean of $X$ equaled 0 and -3 when the mean of $X$ equaled 2. For the normal mixture distribution, $X$ was drawn from $N(-0.875, 0.234)$ and $N(0.875, 0.234)$ with equal probability to have mean 0 and $N(1.125, 0.234)$ and $N(2.875, 0.234)$ with equal probability to have mean 2.

We generated a sample of size $n = 100$ and repeated 1000 simulations for each missingness scenario. For each simulation scenario, 30 percent missingness was induced jointly in $X$ and $X^2$ for five missingness mechanisms: MCAR, MARleft, MARmid, MARtail, and

MARright. Specifically, MCAR denotes that the probability of $X$ being missing is the same for all cases. While with a left-tailed (MARleft), centered (MARmid), both tailed (MARtail) or right-tailed (MARright) missingness mechanism, a higher probability of $X$ being missing is assigned to the units with low, centered, extreme and high values of $Y$ respectively. All missingness was generated with the **ampute** function [Schouten et al., 2018] from the package `MICE` [van Buuren and Groothuis-Oudshoorn, 2011] in `R` [R Core Team, 2021]. The **mice.impute.quadratic** function in the package `MICE` was modified by including data augmentation.

### 2.3.2 Simulation results

We compared five approaches: TTI, ITT, OPC, MPC and SMC-FCS and focused on some exciting findings of OPC, MPC and SMC-FCS. The results of the TTI and ITT simulations reiterated the corresponding conclusions in Table 2.1. In general, TTI did not preserve the quadratic relation. Even it gave unbiased and confidence-valid estimates in some cases (e.g., with MCAR and standard normal distribution $X$). Furthermore, ITT had considerable bias under nearly all combinations of missingness mechanisms and distributions of $X$.

Table 2.3 shows the average biases, the coverage of the nominal 95% confidence intervals, and the average width of confidence intervals for $\beta_1$ and $\beta_2$ when $E(X)$ equals 0. The outcome $Y$ follows a U-shape. With MCAR, MARleft and MARmid and when $X$ is distributed as normal, skewed normal or a mixture of two normals, OPC and MPC gave unbiased estimates and correct CI coverage. The CI coverage of SMC-FCS was close to 95%. However, with $X$ skew-normal distributed MCAR and MARmid, SMC-FCS was slightly biased. With MARtail and MARright, SMC-FCS outperformed OPC and MPC when $X$ followed a normal distribution or a normal mixture distribution. OPC and MPC had slight bias and somewhat reduced CI coverage (approximately 85%) with $X$ distributed according to a normal, a skewed normal or a mixture of two normals. SMC-FCS was unbiased and had CI coverage close to 95% with normal and mixture normal. However, with skewed normal $X$, SMC-FCS was somewhat biased and the CI had slightly lower than nominal coverage.

Table 2.4 demonstrates the mean biases of $\beta_1$ and $\beta_2$, the empirical coverage and the mean width of the corresponding 95% CIs where $X$ is location 2 and scale 1. The observed values of $Y$ are almost on the right arm of the quadratic function. With normal $X$, SMC-FCS consistently yielded confidence-valid estimates because of the congeniality of the analysis model and imputation model. However, with MAR (MARleft, MARmid, MARtial and MARright), SMC-FCS gave a slight bias estimate for $\beta_1$. OPC and MPC gave unbiased results and the CI had approximately 95% coverage with MCAR and MARmid. With MARleft, OPC and MPC were slightly biased for $\beta_1$, but the 95% CI for $\beta_1$ and $\beta_2$ had the correct coverage. With MARtial and MARright, OPC and MPC were unbiased. The CI of MPC (around 90%) had higher nominal coverage than OPC (around 80%). With $X$ distributed according to a skewed normal, OPC and MPC yielded unbiased estimates with MCAR, MARmid, MARtail and MARright but slightly biased estimates with MARleft. The CI coverage of OPC and MPC was close to 95% with MCAR, MARleft and MARmid. Like the case with normal $X$, the CI of MPC (around 90%) had better coverage than OPC (around 85%). With $X$ distributed as a mixture of two normals, SMC-FCS gave biased results under all missingness mechanisms and its CI had somewhat reduced coverage with MARright. OPC and MPC were unbiased with MCAR, MARmid and MARtail but biased with MARleft and MARright. MPC had CI coverage close to 95% under all missingness mechanisms. The CI from OPC had correct coverage with MCAR, MARleft and MARmid, but approximately 85% coverage with MARtail and MARright.

We investigated if the biases of SMC-FCS were caused by Monte Carlo error. Figures 2.2(a) and 2.2(b) demonstrate that the bias is somewhat systematic and not due to simulation error. The estimates for $\beta_1$ and $\beta_2$ show primarily overestimation and underestimation, respectively. This implies that, when applying SMC-FCS, the explicit specification of the distribution of the incomplete variable with the quadratic effect may need careful consideration.

## 2.4 Conclusion

We evaluated the performance of four imputation approaches for incomplete data problems where the model of scientific interest contains squared terms. We improved the performance of the polynomial combination method by incorporating a data augmentation step, thus realizing more plausible imputations when the missingness covariate relates almost exclusively to one arm of the quadratic curve.

In our simulation studies, ITT gave biased estimates under almost all combinations of experimental factors, and TTI cannot preserve the quadratic relations in the imputed data. With normally distributed predictors and right-tailed missingness mechanisms, the performance of SMC-FCS was superior to that of MPC, with coverages closer to the nominal level. However, when the normality assumption is violated, the polynomial combination yields less biased estimates. Overall, both the SMC-FCS and polynomial combination methods produce plausible imputations of squared terms and outperform TTI and ITT. Differences between the approaches only become apparent under intense MARtail and MARright scenarios in simulation. However, these two mechanisms are more extreme than we are likely to see in practice since there is a strong relationship between the outcome $Y$ and the probability of the variable $X$ being unobserved in the tail. All in all, when differences in performance are found, such differences are small, and it may be challenging to interpret them as meaningful. This means that, in practice, the choice for an imputation approach could largely be a choice of preference.

If there is a solid, well-known scientific model, we highly recommend using SMC-FCS to sharpen results. The substantive model would then be correctly specified, ensuring that the distribution from which imputations are generated is compatible. SMC-FCS is a reliable model-based method to impute predictors with quadratic effects. It is theoretically well-grounded, and procedures are available for substantive models based on standard regression, discrete outcomes and proportional hazards [van Buuren, 2018]. However, with an increasing number of variables in the dataset, it becomes increasingly challenging to infer the correct substantive model based on the incomplete data a priori. The strategy of applying SMC-FCS in practice is performing model selection once imputed datasets are generated to ensure the accuracy of substantive model specification, which is not a trivial process [Bartlett et al., 2015]. Usually, the substantive model is specified according to prior studies or assumptions.

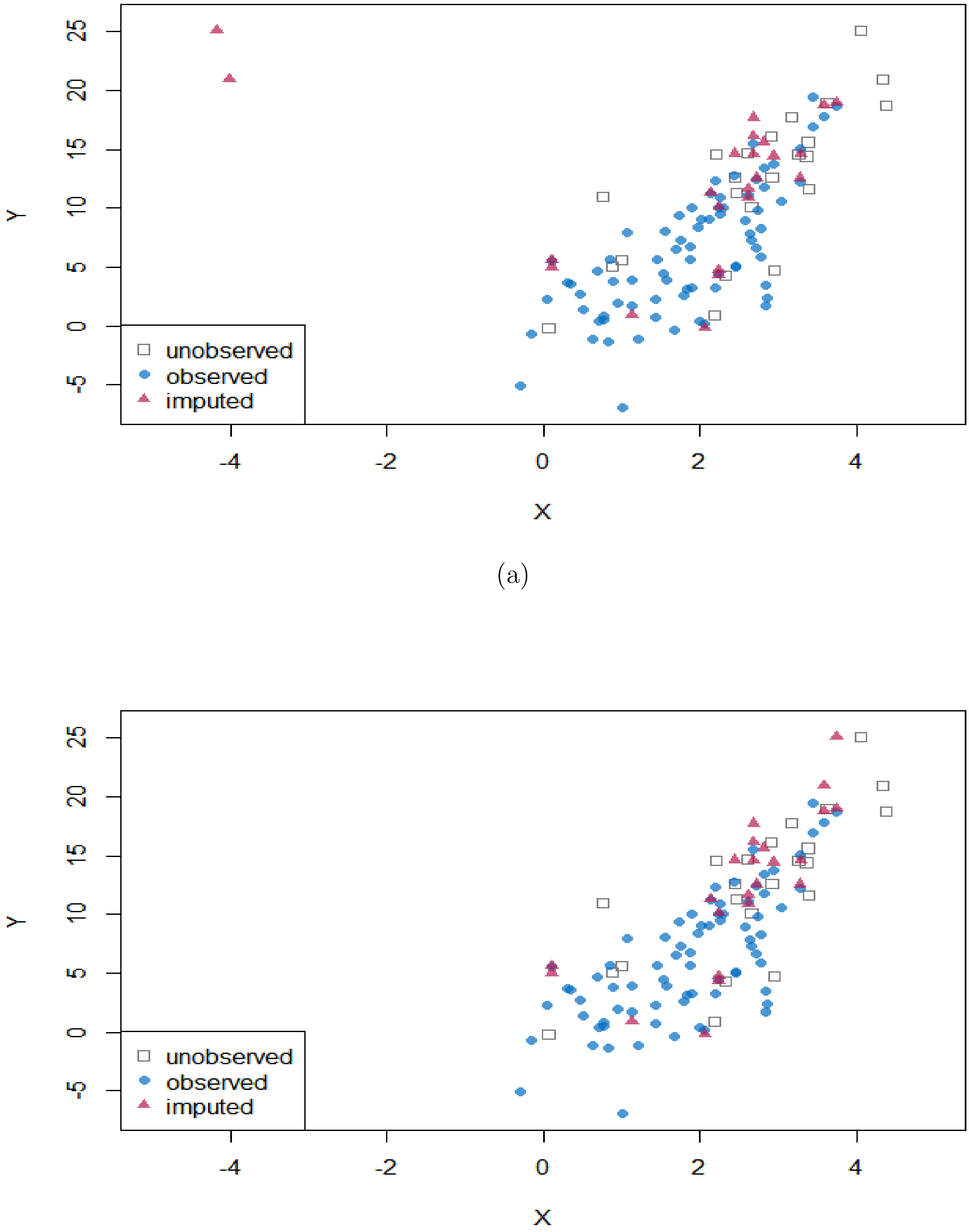
In contrast, we advise using the polynomial combination approach when the scientific model is less specific or when modeling efforts are challenging. It is proven to be a valid data-driven imputation method that is flexible in applying because we only need to specify the quadratic term. This makes it straightforward to implement in any imputation effort. The polynomial combination method is based on predictive mean matching, and the performance of imputation procedures involving PMM are proven to work well in a wide range of research problems [Vink et al., 2015, Rubin, 1986, Little, 1988]. Therefore, we expect that the polynomial combination approach could be of great practical importance in incomplete data analyses with squared terms.

| | Missingness Mechanism | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MCAR | | | MARleft | | | MARmid | | | MARtail | | | MARright | | |
| | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw |
| Normal | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.01 | 0.94 | 0.52 | -0.05 | 0.92 | 0.47 | -0.01 | 0.96 | 0.51 | 0.11 | 0.88 | 0.63 | 0.21 | 0.78 | 0.74 |
| $\beta_2$ | 0 | 0.94 | 0.39 | -0.06 | 0.9 | 0.34 | -0.03 | 0.93 | 0.36 | 0.1 | 0.86 | 0.51 | 0.2 | 0.73 | 0.62 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.05 | 0.9 | 0.9 | -0.02 | 0.98 | 0.94 | -0.09 | 0.98 | 0.86 | -0.04 | 0.99 | 1.09 | -0.04 | 0.98 | 1.13 |
| $\beta_2$ | -0.25 | 0.88 | 0.89 | -0.25 | 0.86 | 0.86 | -0.2 | 0.9 | 0.78 | -0.34 | 0.84 | 1.17 | -0.31 | 0.87 | 1.16 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.03 | 0.94 | 0.56 | 0.01 | 0.95 | 0.46 | 0.01 | 0.95 | 0.48 | 0.08 | 0.87 | 0.75 | 0.09 | 0.88 | 0.87 |
| $\beta_2$ | 0.03 | 0.92 | 0.41 | 0 | 0.94 | 0.33 | 0 | 0.94 | 0.34 | 0.13 | 0.84 | 0.6 | 0.15 | 0.83 | 0.72 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.03 | 0.94 | 0.56 | 0.01 | 0.95 | 0.46 | 0.01 | 0.95 | 0.47 | 0.1 | 0.86 | 0.75 | 0.1 | 0.88 | 0.86 |
| $\beta_2$ | 0.03 | 0.92 | 0.41 | 0 | 0.94 | 0.33 | 0 | 0.94 | 0.34 | 0.13 | 0.84 | 0.6 | 0.15 | 0.83 | 0.72 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0 | 0.96 | 0.53 | -0.01 | 0.95 | 0.47 | 0.01 | 0.94 | 0.49 | 0 | 0.96 | 0.6 | 0 | 0.95 | 0.67 |
| $\beta_2$ | 0 | 0.95 | 0.39 | 0 | 0.95 | 0.33 | 0 | 0.95 | 0.34 | 0.01 | 0.95 | 0.51 | 0.02 | 0.95 | 0.57 |
| Skewed-normal | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.15 | 0.95 | 3.21 | -0.25 | 0.92 | 2.98 | 0.15 | 0.95 | 2.79 | -0.43 | 0.92 | 4.93 | -0.26 | 0.94 | 5.75 |
| $\beta_2$ | 0.09 | 0.94 | 1.48 | 0.04 | 0.93 | 1.28 | -0.09 | 0.94 | 1.23 | 0.35 | 0.88 | 2.59 | 0.39 | 0.91 | 3.28 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.3 | 0.94 | 2.55 | 0.27 | 0.93 | 2.24 | 0.18 | 0.96 | 2.39 | 0.5 | 0.9 | 2.91 | 0.41 | 0.94 | 3.18 |
| $\beta_2$ | -0.13 | 0.96 | 1.27 | -0.12 | 0.93 | 1.01 | -0.07 | 0.96 | 1.07 | -0.24 | 0.9 | 1.71 | -0.15 | 0.95 | 1.91 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0 | 0.93 | 2.68 | -0.02 | 0.94 | 2.49 | 0.03 | 0.93 | 2.4 | -0.07 | 0.86 | 3.64 | 0.01 | 0.82 | 3.79 |
| $\beta_2$ | 0.03 | 0.92 | 1.27 | 0.01 | 0.94 | 1.11 | -0.01 | 0.94 | 1.08 | 0.16 | 0.85 | 1.98 | 0.13 | 0.82 | 2.16 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.01 | 0.95 | 2.69 | -0.02 | 0.94 | 2.47 | 0.03 | 0.93 | 2.39 | -0.02 | 0.92 | 3.87 | 0.03 | 0.88 | 4 |
| $\beta_2$ | 0.03 | 0.94 | 1.29 | 0.01 | 0.94 | 1.11 | -0.01 | 0.93 | 1.08 | 0.15 | 0.91 | 2.09 | 0.15 | 0.86 | 2.26 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.14 | 0.94 | 2.59 | -0.01 | 0.95 | 2.39 | -0.15 | 0.96 | 2.43 | -0.41 | 0.93 | 3.51 | -0.49 | 0.91 | 3.56 |
| $\beta_2$ | 0.09 | 0.95 | 1.21 | 0 | 0.94 | 1.08 | 0.07 | 0.96 | 1.07 | 0.28 | 0.91 | 1.87 | 0.39 | 0.88 | 1.97 |
| Normal-mixture | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0 | 0.96 | 0.4 | -0.05 | 0.93 | 0.38 | 0 | 0.95 | 0.4 | 0.02 | 0.95 | 0.43 | 0.1 | 0.86 | 0.46 |
| $\beta_2$ | 0 | 0.96 | 0.46 | -0.05 | 0.94 | 0.42 | -0.04 | 0.95 | 0.44 | 0.02 | 0.96 | 0.52 | 0.09 | 0.91 | 0.57 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.04 | 0.98 | 0.62 | 0.03 | 0.98 | 0.59 | -0.02 | 0.98 | 0.58 | -0.07 | 0.98 | 0.65 | -0.09 | 0.98 | 0.68 |
| $\beta_2$ | -0.42 | 0.59 | 0.98 | -0.36 | 0.71 | 0.96 | -0.32 | 0.72 | 0.86 | -0.55 | 0.35 | 0.97 | -0.52 | 0.4 | 0.96 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.02 | 0.94 | 0.38 | 0 | 0.93 | 0.36 | 0.01 | 0.94 | 0.37 | 0.06 | 0.88 | 0.44 | 0.08 | 0.87 | 0.48 |
| $\beta_2$ | 0.01 | 0.94 | 0.44 | 0 | 0.93 | 0.41 | 0 | 0.94 | 0.41 | 0.04 | 0.92 | 0.53 | 0.05 | 0.92 | 0.58 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.02 | 0.94 | 0.38 | 0 | 0.93 | 0.36 | 0.01 | 0.94 | 0.36 | 0.06 | 0.89 | 0.44 | 0.08 | 0.87 | 0.48 |
| $\beta_2$ | 0.01 | 0.95 | 0.44 | 0 | 0.94 | 0.41 | 0 | 0.94 | 0.4 | 0.04 | 0.92 | 0.53 | 0.05 | 0.93 | 0.58 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.01 | 0.95 | 0.4 | -0.02 | 0.95 | 0.38 | 0.02 | 0.95 | 0.38 | -0.05 | 0.93 | 0.43 | -0.05 | 0.93 | 0.48 |
| $\beta_2$ | -0.02 | 0.95 | 0.46 | 0.01 | 0.94 | 0.41 | -0.03 | 0.94 | 0.42 | -0.03 | 0.94 | 0.51 | -0.06 | 0.93 | 0.57 |

Table 2.3: Simulation results for five missingness mechanisms when imputing a squared term regression where the mean of X equals 0. Shown are absolute bias of the estimate, coverage of the 95% confidence interval for the estimate and the average confidence interval width.

| | Missingness Mechanism | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MCAR | | | MARleft | | | MARmid | | | MARtail | | | MARright | | |
| | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw | Bias | Cov | Ciw |
| Normal | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.58 | 0.93 | 7.17 | -1 | 0.92 | 8 | -0.08 | 0.95 | 6.04 | -1.32 | 0.91 | 10.11 | -0.89 | 0.93 | 9.24 |
| $\beta_2$ | 0.11 | 0.93 | 1.66 | 0.15 | 0.93 | 1.73 | 0 | 0.95 | 1.39 | 0.31 | 0.89 | 2.43 | 0.29 | 0.91 | 2.4 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.71 | 0.94 | 5.61 | 0.6 | 0.94 | 5.57 | 0.5 | 0.96 | 5.27 | 0.99 | 0.92 | 5.95 | 1.02 | 0.93 | 4.97 |
| $\beta_2$ | -0.19 | 0.95 | 1.41 | -0.14 | 0.95 | 1.31 | -0.13 | 0.96 | 1.28 | -0.28 | 0.9 | 1.58 | -0.3 | 0.91 | 1.63 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.05 | 0.92 | 5.28 | -0.15 | 0.93 | 5.56 | 0.01 | 0.93 | 4.98 | -0.12 | 0.87 | 6.04 | 0.04 | 0.8 | 5.65 |
| $\beta_2$ | 0.02 | 0.93 | 1.27 | 0.03 | 0.94 | 1.27 | 0 | 0.94 | 1.17 | 0.05 | 0.87 | 1.51 | 0.03 | 0.82 | 1.48 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.05 | 0.93 | 5.17 | -0.15 | 0.94 | 5.39 | 0 | 0.91 | 4.77 | -0.03 | 0.93 | 6.26 | 0.05 | 0.89 | 6.02 |
| $\beta_2$ | 0.02 | 0.93 | 1.25 | 0.03 | 0.94 | 1.24 | 0.01 | 0.92 | 1.13 | 0.04 | 0.91 | 1.56 | 0.04 | 0.88 | 1.57 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.03 | 0.95 | 5.44 | -0.11 | 0.95 | 5.78 | -0.06 | 0.95 | 5.14 | -0.22 | 0.94 | 6.02 | -0.1 | 0.97 | 5.69 |
| $\beta_2$ | 0 | 0.95 | 1.3 | 0.03 | 0.94 | 1.31 | 0.01 | 0.95 | 1.25 | 0.06 | 0.94 | 1.5 | 0.04 | 0.96 | 1.48 |
| Skewed-normal | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.15 | 0.93 | 2.95 | -0.45 | 0.92 | 4.14 | -0.06 | 0.95 | 2.56 | -0.32 | 0.95 | 4.07 | -0.07 | 0.94 | 2.97 |
| $\beta_2$ | 0.06 | 0.93 | 1.27 | 0.14 | 0.92 | 1.59 | 0.03 | 0.95 | 1.11 | 0.12 | 0.94 | 1.73 | 0.06 | 0.93 | 1.41 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.46 | 0.92 | 2.64 | 0.29 | 0.93 | 2.86 | 0.35 | 0.94 | 2.49 | 0.62 | 0.87 | 2.83 | 0.73 | 0.81 | 2.48 |
| $\beta_2$ | -0.28 | 0.88 | 1.26 | -0.13 | 0.94 | 1.23 | -0.21 | 0.9 | 1.16 | -0.37 | 0.8 | 1.37 | -0.48 | 0.64 | 1.26 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.01 | 0.92 | 2.24 | -0.11 | 0.93 | 2.74 | 0.02 | 0.93 | 2.19 | -0.07 | 0.88 | 2.5 | 0 | 0.85 | 2.19 |
| $\beta_2$ | 0 | 0.93 | 0.99 | 0.05 | 0.93 | 1.14 | -0.01 | 0.94 | 0.96 | 0.04 | 0.89 | 1.11 | 0.02 | 0.86 | 1.03 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0 | 0.92 | 2.18 | -0.11 | 0.93 | 2.66 | 0 | 0.92 | 2.02 | -0.04 | 0.94 | 2.59 | 0.02 | 0.91 | 2.28 |
| $\beta_2$ | 0.01 | 0.93 | 0.96 | 0.05 | 0.93 | 1.11 | 0 | 0.93 | 0.89 | 0.04 | 0.95 | 1.15 | 0.02 | 0.9 | 1.06 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.1 | 0.95 | 2.46 | -0.04 | 0.95 | 3.01 | 0.09 | 0.95 | 2.27 | 0.14 | 0.95 | 2.65 | 0.2 | 0.93 | 2.22 |
| $\beta_2$ | -0.05 | 0.94 | 1.08 | 0.03 | 0.95 | 1.22 | -0.04 | 0.94 | 1 | -0.08 | 0.94 | 1.2 | -0.14 | 0.91 | 1.04 |
| Normal-mixture | | | | | | | | | | | | | | | |
| `TTI` | | | | | | | | | | | | | | | |
| $\beta_1$ | -1.5 | 0.9 | 10.48 | -2.15 | 0.86 | 11.93 | -0.71 | 0.93 | 8.8 | -2.35 | 0.87 | 13.63 | -1.73 | 0.91 | 12.47 |
| $\beta_2$ | 0.33 | 0.9 | 2.52 | 0.43 | 0.87 | 2.73 | 0.16 | 0.94 | 2.11 | 0.51 | 0.88 | 3.29 | 0.43 | 0.9 | 3.15 |
| `ITT` | | | | | | | | | | | | | | | |
| $\beta_1$ | 1.38 | 0.88 | 6.4 | 0.83 | 0.92 | 6.13 | 1.07 | 0.92 | 6.33 | 1.76 | 0.8 | 6.08 | 2.08 | 0.78 | 6.35 |
| $\beta_2$ | -0.35 | 0.88 | 1.62 | -0.19 | 0.94 | 1.52 | -0.25 | 0.92 | 1.58 | -0.49 | 0.76 | 1.57 | -0.57 | 0.72 | 1.63 |
| `OPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.05 | 0.92 | 6.28 | -0.24 | 0.92 | 6.51 | 0.02 | 0.93 | 6.05 | -0.09 | 0.88 | 6.61 | 0.19 | 0.85 | 6.43 |
| $\beta_2$ | 0.02 | 0.92 | 1.53 | 0.06 | 0.92 | 1.57 | 0 | 0.94 | 1.47 | 0.04 | 0.87 | 1.64 | -0.03 | 0.86 | 1.62 |
| `MPC` | | | | | | | | | | | | | | | |
| $\beta_1$ | -0.04 | 0.93 | 6.24 | -0.23 | 0.92 | 6.39 | 0.02 | 0.91 | 5.92 | -0.04 | 0.93 | 6.77 | 0.24 | 0.92 | 6.69 |
| $\beta_2$ | 0.02 | 0.93 | 1.53 | 0.05 | 0.93 | 1.54 | 0 | 0.92 | 1.45 | 0.03 | 0.92 | 1.68 | -0.03 | 0.92 | 1.69 |
| `SMC-FCS` | | | | | | | | | | | | | | | |
| $\beta_1$ | 0.32 | 0.95 | 6.37 | -0.32 | 0.96 | 6.85 | 0.25 | 0.94 | 6.18 | 0.5 | 0.94 | 6.58 | 0.95 | 0.91 | 6.69 |
| $\beta_2$ | -0.07 | 0.94 | 1.57 | 0.09 | 0.95 | 1.65 | -0.04 | 0.94 | 1.5 | -0.14 | 0.92 | 1.66 | -0.25 | 0.9 | 1.71 |

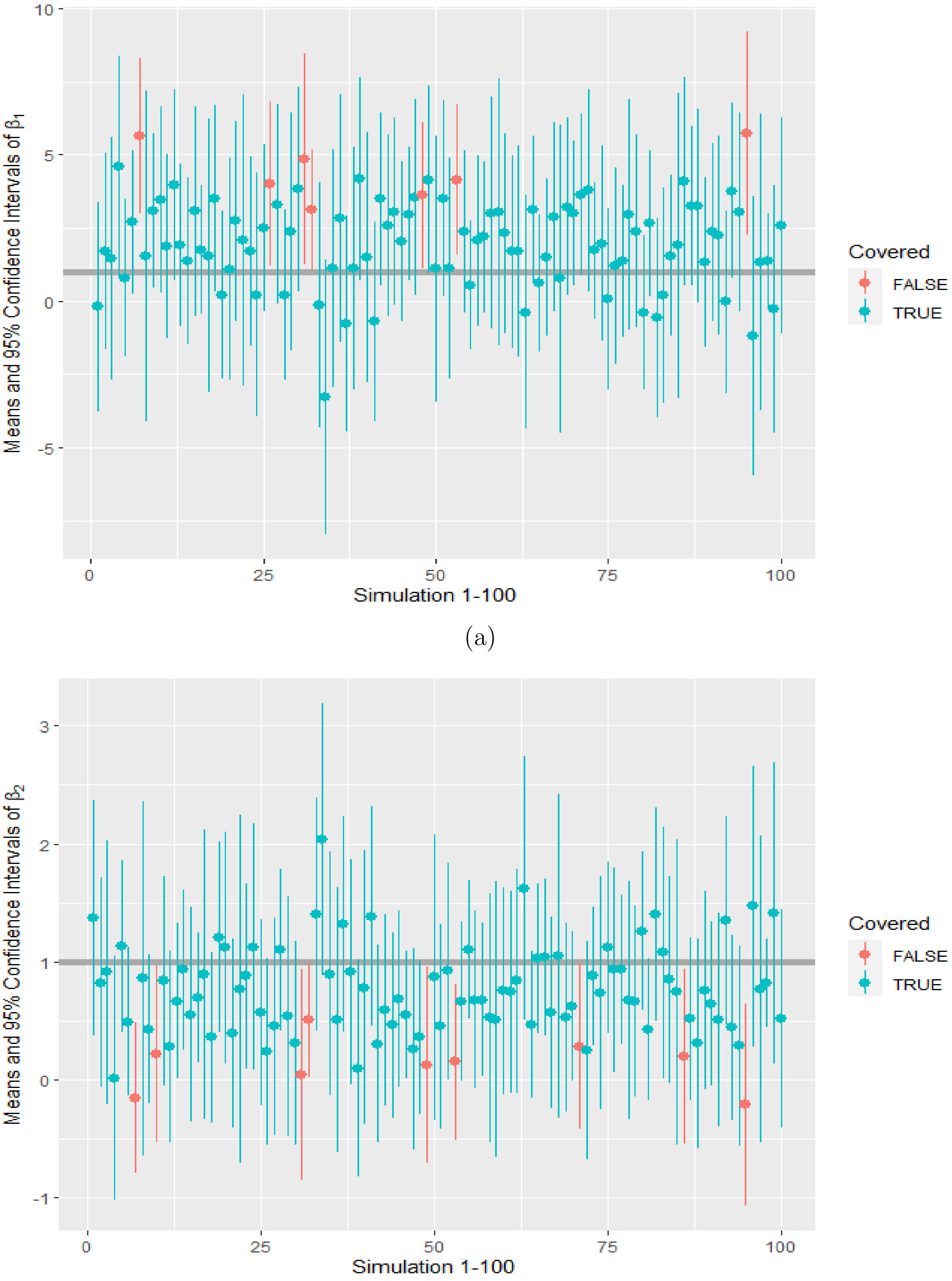
Table 2.4: Simulation results for five missingness mechanisms when imputing a squared term regression where the mean of X equals 2. Shown are absolute bias of the estimate, coverage of the 95% confidence interval for the estimate and the average confidence interval width.



(a)

(b)

Figure 2.1: Imputations (triangles) generated by OPC and MPC. We see that in (a) some imputations fall outside of the range of the observed (circle) and unobserved values (square), due to the OPC algorithm assigning the donor values to the incorrect distinct real root. In (b) the MPC approach assigns the imputations to the distinct real root that corresponds to the observed and unobserved data.



(a)

(b)

Figure 2.2: The plot of means and confidence intervals of $\beta_1$ and $\beta_2$ from the first 100 simulations. The model of interest is $Y = X + X^2 + \epsilon$, where $X \sim \frac{1}{2}N(1.125, 0.234) + \frac{1}{2}N(2.875, 0.234)$. The missingness mechanism is MARright and the imputation approach is SMC-FCS. The imputations seem to primarily overestimate the true parameter estimate of $\beta_1$ in (a) and underestimate the true parameter estimate of $\beta_2$ in (b).

# Chapter 3

# Generalizing univariate predictive mean matching to impute multiple variables simultaneously

## 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.

## 3.1 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 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 [Rubin, 1987, p.76].

Two widely used strategies for imputing multivariate missing data are joint modeling (JM) and fully conditional specification (FCS). Joint modeling was proposed by Rubin [1987] and especially developed by Schafer [1997]. 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 [2007], follows an iterative scheme that imputes each incomplete variable based on a conditionally specified model. Fully conditional specification allows for tremendous flexibility in multivariate model design and flexibility in imputing non-normal variables, especially discrete variables [Goldstein et al., 2014]. However, FCS may suffer from incompatibility problems, and computational shortcuts like the sweep operator cannot be applied to facilitate computation [van Buuren, 2018]. 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 [2018, section 4.7.2] 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 [Little, 1988]. 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 [van Buuren and Groothuis-Oudshoorn, 2011]. We tailor PMM to the block imputation framework, which will widen its application. More computational details and properties of PMM would be addressed in section 3.2.

For a comprehensive overview of missing data analysis, we refer to Little and Rubin [2019] for a comparison of approaches to missing data other than multiple imputation (e.g, ad-hoc methods, maximum likelihood estimation and weighting methods). Schafer [1999], Sinharay et al. [2001] and Allison [2001] introduced basic concepts and general methods of MI. Schafer and Graham [2002] discussed practical issues of application of MI. Various sophisticated missing data analysis were developed on the fields of multilevel model [Longford, 2001], structural equation modeling [Olinsky et al., 2003, Allison, 2003], longitudinal data analysis [Twisk and de Vente, 2002, Demirtas, 2004] and meta-analysis [Pigott, 2001]. Schafer [2003] compared Bayesian MI methods with maximum likelihood estimation. Seaman and White [2013] gave an overview of the use of inverse probability weighting in missing data problems. Ibrahim et al. [2005] 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`) [van Buuren and Groothuis-Oudshoorn, 2011].

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 3.3 provides a simple comparison between PMM and MPMM. Section 3.4 is a simulation study investigating whether MPMM yields valid estimates and preserves functional relations between imputed values. The discussion closes the paper.

## 3.2 Multivariate Predictive Mean Matching

### 3.2.1 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 [Israels, 1987]. The basic idea of canonical regression analysis is quite old and has been discussed under different names, such as Rank-reduced regression [Izenman, 1975] and partial least squares [Sun et al., 2009].

Let us consider the equation

$$\boldsymbol{\alpha' Y} = \boldsymbol{\beta X} + \epsilon. \tag{3.1}$$

We aim to minimize the variance $\epsilon$ with respect to $\alpha$ and $\beta$ 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

$$R^2_{\alpha' y.x} = \frac{\boldsymbol{\alpha' \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \alpha}}{\boldsymbol{\alpha' \Sigma_{yy} \alpha}}, \tag{3.2}$$

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 [1968], 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.

### 3.2.2 Predictive mean matching (PMM)

PMM was first proposed by Rubin [1986]and formalized by Little [1988]. 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 [van Buuren, 2018].

## Computational details

We elaborate the algorithm of predictive mean matching for the clear illustration of its merger with canonical regression analysis [Vink et al., 2015]. $\boldsymbol{X_{obs}}$, 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 $\chi^2$ 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 [van Buuren and Groothuis-Oudshoorn, 2011, Vink et al., 2015, Heitjan and Little, 1991, Morris et al., 2014, Vink et al., 2014]. 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 [van Buuren, 2018]. For the same reason, PMM yields acceptable imputations even when normality assumptions are violated [Vink et al., 2014]. 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 [James R. Carpenter, 2013].

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. [2012] 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.

### 3.2.3 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

$$\begin{pmatrix} \boldsymbol{\Sigma_{y_{obs}y_{obs}}} & \boldsymbol{\Sigma_{y_{obs}x_{obs}}} \\ \boldsymbol{\Sigma_{x_{obs}y_{obs}}} & \boldsymbol{\Sigma_{x_{obs}x_{obs}}} \end{pmatrix}$$

2. Find the largest eigenvalue $\lambda^2$ of $\boldsymbol{\Sigma^{-1}_{y_{obs}y_{obs}}\Sigma_{y_{obs}x_{obs}}\Sigma^{-1}_{x_{obs}x_{obs}}\Sigma_{x_{obs}y_{obs}}}$ and its corresponding right-hand eigenvector $\alpha$

3. Calculate the linear combination $\boldsymbol{\alpha' Y}$ for all completely observed individuals in the

sample: $\boldsymbol{V_{obs}} = \boldsymbol{\alpha' Y_{obs}}$

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 $\chi^2$ variate with $N_{obs} - j$ degrees of freedom

6. Draw $\beta^*$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$and covariance matrix $\boldsymbol{\sigma^{2*}(X_{obs}^{\mathsf{T}}X_{obs})^{-1}}$

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) [Rencher, 2003, chapter 10] and redundancy analysis (RA) [Van Den Wollenberg, 1977]. 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.

## 3.3 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.

### 3.3.1 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 $\rho$ 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 (MAR-right) 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 $\alpha_0$ was chosen to make the probability of jointly missing $\mathbf{Y}$ equal to 0.4. Missing values were induced with the `ampute` function [Schouten et al., 2018] from the package `MICE` [van Buuren and Groothuis-Oudshoorn, 2011] in `R` [R Core Team, 2021]. The correlation $\rho$ 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 [Rubin, 1987, p.76].

### 3.3.2 Results

| | | $E(Y_1)$ | | | | $\rho(Y_1, Y_2)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PMM | | PMM-CRA | | PMM | | PMM-CRA | |
| $\rho$ | scenario | 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 3.1: Simulation results for evaluating whether MPMM provide valid imputations at the univariate level.

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 MARright 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.

## 3.4 Evaluation

To investigate the performance of MPMM when there are relations in the incomplete data, we performed the following simulation studies.

### 3.4.1 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

$$\boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^{\mathbf{2}} + \epsilon \tag{3.3}$$

where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, both predictor $\boldsymbol{X}$ and error term $\epsilon$ 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 $\boldsymbol{X}^{\mathbf{2}}$ 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 $\boldsymbol{X}^{\mathbf{2}}$ 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 [2009]. We also apply polynomial combination proposed by Vink and van Buuren [2013]. PC imputes the combination of $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathbf{2}}$ by predicted mean matching and then decomposes it by solving a quadratic equation for $\boldsymbol{X}$. The polynomial

[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.

combination method is implemented by `mice.impute.quadratic` function in the `R MICE` package. Finally, SMC-FCS is proposed by Bartlett et al. [2015]. In general, it imputes the missing variable based on the formula:

$$\begin{aligned} f(\boldsymbol{X_i}|\boldsymbol{X_{-i}},\boldsymbol{Y}) &= \tfrac{f(\boldsymbol{X_i},\boldsymbol{X_{-i}},\boldsymbol{Y})}{f(\boldsymbol{Y},\boldsymbol{X_{-i}})} \\ &\propto f(\boldsymbol{Y}|\boldsymbol{X_i},\boldsymbol{X_{-i}})f(\boldsymbol{X_i}|\boldsymbol{X_{-i}}). \end{aligned} \tag{3.4}$$

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.

### Results

Table 3.2 displays the results of the simulation, including estimates of $\alpha$, $\beta_1$, $\beta_2$, $\sigma_\epsilon$, $R^2$ and the coverage of nominal 95% confidence intervals of $\beta_1$ and $\beta_2$. 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 $\beta_1$ and $\beta_2$ 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 [de Jong et al., 2014]. 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.

|  | Missingness Mechanism |  |  |  |  |
|---|---|---|---|---|---|
|  | MCAR | MARleft | MARmid | MARtail | MARright |
| *Transform, then impute* |  |  |  |  |  |
| Intercept ($\alpha$) | 0 | 0.15 | -0.04 | 0 | -0.11 |
| Slope of $X$ ($\beta_1$) | 1(0.93) | 0.93(0.02) | 0.97(0.68) | 1.13(0) | 1.27(0) |
| Slope of $X^2$ ($\beta_2$) | 1(0.92) | 0.93(0) | 0.96(0.13) | 1.13(0) | 1.27(0) |
| Residual SD ($\sigma_\epsilon$) | 1 | 0.96 | 1 | 1.06 | 1.13 |
| $R^2$ | 0.75 | 0.77 | 0.75 | 0.72 | 0.68 |
| *Impute, then transform* |  |  |  |  |  |
| Intercept ($\alpha$) | 0.32 | 0.22 | 0.2 | 0.45 | 0.49 |
| Slope of $X$ ($\beta_1$) | 0.94(0.62) | 0.97(0.91) | 0.89(0.08) | 1(0.99) | 1.04(0.92) |
| Slope of $X^2$ ($\beta_2$) | 0.68(0) | 0.68(0) | 0.74(0) | 0.62(0) | 0.7(0) |
| Residual SD ($\sigma_\epsilon$) | 1.41 | 1.36 | 1.35 | 1.52 | 1.57 |
| $R^2$ | 0.5 | 0.54 | 0.55 | 0.42 | 0.38 |
| *PC* |  |  |  |  |  |
| Intercept ($\alpha$) | 0 | 0 | 0 | -0.05 | -0.06 |
| Slope of $X$ ($\beta_1$) | 1(0.93) | 1(0.93) | 1(0.93) | 1(0.85) | 1(0.82) |
| Slope of $X^2$ ($\beta_2$) | 1.01(0.9) | 1(0.94) | 1(0.93) | 1.07(0.12) | 1.09(0.09) |
| Residual SD ($\sigma_\epsilon$) | 1 | 1 | 1 | 1.05 | 1.07 |
| $R^2$ | 0.75 | 0.75 | 0.75 | 0.72 | 0.71 |
| *PMM-CRA* |  |  |  |  |  |
| Intercept ($\alpha$) | 0 | 0 | 0 | -0.03 | -0.03 |
| Slope of $X$ ($\beta_1$) | 1(0.93) | 1(0.93) | 1(0.91) | 1.04(0.47) | 1.06(0.4) |
| Slope of $X^2$ ($\beta_2$) | 1(0.91) | 1(0.95) | 1(0.93) | 1.05(0.25) | 1.07(0.23) |
| Residual SD ($\sigma_\epsilon$) | 1 | 1 | 1 | 1.05 | 1.07 |
| $R^2$ | 0.75 | 0.75 | 0.75 | 0.72 | 0.71 |
| *SMC-FCS* |  |  |  |  |  |
| Intercept ($\alpha$) | 0.01 | 0 | 0 | 0.03 | 0.05 |
| Slope of $X$ ($\beta_1$) | 1(0.96) | 1(0.95) | 1(0.95) | 1(0.97) | 1.01(0.97) |
| Slope of $X^2$ ($\beta_2$) | 1(0.95) | 1(0.96) | 1(0.94) | 1(0.96) | 1.01(0.93) |
| Residual SD ($\sigma_\epsilon$) | 1.04 | 1 | 1 | 1.11 | 1.12 |
| $R^2$ | 0.73 | 0.75 | 0.75 | 0.69 | 0.68 |

Table 3.2: 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 $R^2 = .75$.

Although *'impute, then transform'* method preserved the squared relationship, it resulted in severely biased estimates, even with MCAR. The CI coverage of $\beta_2$ was considerably poor, with all cases of missingness mechanisms. With MCAR, *'transform, then impute'* method yielded unbiased regression estimates and correct CI coverage for $\beta_1$ and

$\beta_2$. However, TTI distorted the quadratic relation between $\boldsymbol{X}$ and $\boldsymbol{X^2}$. It also gave severely biased results, and the CIs for $\beta_1$ and $\beta_2$ 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 $\sigma_\epsilon$ 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 3.1 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.

### 3.4.2 Linear regression with interaction term

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

#### Simulation conditions

The dependent variable $Y$ was generated according to the analysis model

$$\boldsymbol{Y} = \alpha + \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \beta_3\boldsymbol{X_1}\boldsymbol{X_2} + \epsilon \tag{3.5}$$

where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$, two predictors $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and error term $\epsilon$ 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.

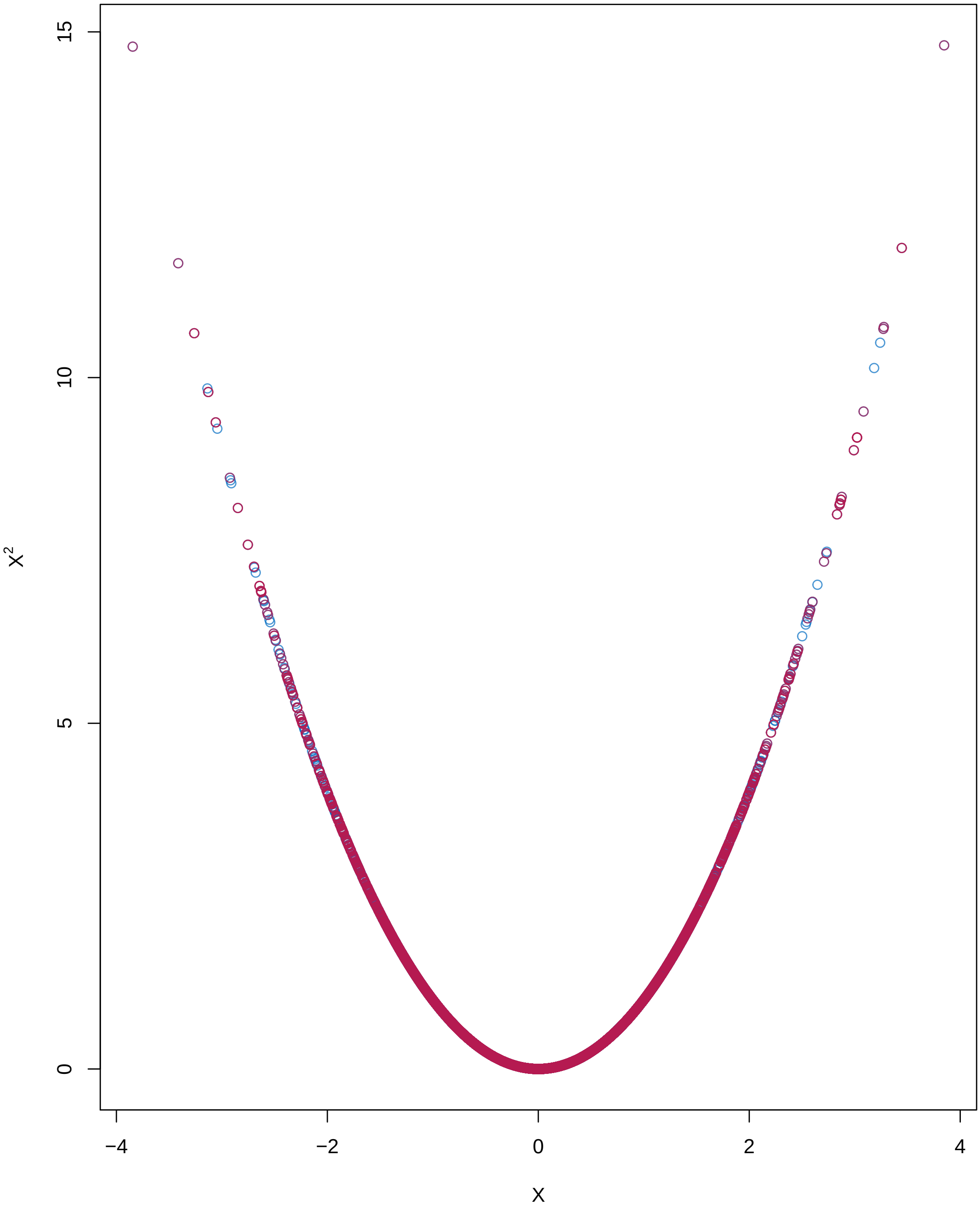


Figure 3.1: Predictive mean matching based on canonical regression analysis. Observed (blue) and imputed values (red) for $X$ and $X^2$.

**Estimation methods**

We evaluated and compared the same methods as under section 3.4.1, except the polynomial combination method. The model-based imputation method ensures a compatible imputation model by accommodating the designed model $\boldsymbol{Y} = \boldsymbol{X_1} + \boldsymbol{X_2} + \boldsymbol{X_1}\boldsymbol{X_2} + \epsilon$.

### Results

| | Missingness Mechanism | | | | |
|---|---|---|---|---|---|
| | MCAR | MARleft | MARmid | MARtail | MARright |
| *Transform, then impute* | | | | | |
| Intercept ($\alpha$) | 0 | 0.05 | -0.05 | 0.06 | 0.05 |
| Slope of $X_1$ ($\beta_1$) | 1(0.93) | 0.96(0.4) | 1(0.94) | 1.05(0.42) | 1.08(0.05) |
| Slope of $X_2$ ($\beta_2$) | 1(0.94) | 0.96(0.4) | 1(0.96) | 1.05(0.38) | 1.09(0.02) |
| Slope of $X_1X_2(\beta_3)$ | 1(0.94) | 0.96(0.53) | 0.95(0.25) | 1.06(0.31) | 1.09(0.02) |
| Residual SD ($\sigma_\epsilon$) | 1 | 0.97 | 1 | 1.02 | 1.04 |
| $R^2$ | 0.75 | 0.76 | 0.75 | 0.74 | 0.73 |
| *Impute, then transform* | | | | | |
| Intercept ($\alpha$) | 0 | -0.04 | -0.01 | 0.01 | 0.11 |
| Slope of $X_1$ ($\beta_1$) | 0.98(0.88) | 1.05(0.51) | 0.96(0.71) | 0.98(0.9) | 0.95(0.69) |
| Slope of $X_2$ ($\beta_2$) | 0.98(0.88) | 1.05(0.48) | 0.96(0.73) | 0.98(0.92) | 0.95(0.69) |
| Slope of $X_1X_2(\beta_3)$ | 0.64(0) | 0.64(0) | 0.7(0) | 0.54(0) | 0.61(0) |
| Residual SD ($\sigma_\epsilon$) | 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 ($\alpha$) | 0 | 0 | 0 | 0.01 | 0.02 |
| Slope of $X_1$ ($\beta_1$) | 1(0.93) | 1(0.86) | 1(0.92) | 1.02(0.8) | 1.02(0.73) |
| Slope of $X_2$ ($\beta_2$) | 1(0.93) | 1(0.84) | 1(0.93) | 1.02(0.8) | 1.02(0.77) |
| Slope of $X_1X_2(\beta_3)$ | 1(0.94) | 1.01(0.86) | 1(0.93) | 1.02(0.71) | 1.03(0.68) |
| Residual SD ($\sigma_\epsilon$) | 1 | 1.01 | 1 | 1.03 | 1.03 |
| $R^2$ | 0.75 | 0.74 | 0.75 | 0.74 | 0.74 |
| *SMC-FCS* | | | | | |
| Intercept ($\alpha$) | 0 | -0.01 | 0 | 0.01 | 0.03 |
| Slope of $X_1$ ($\beta_1$) | 1(0.95) | 1.01(0.95) | 1(0.95) | 1(0.96) | 0.99(0.95) |
| Slope of $X_2$ ($\beta_2$) | 0.99(0.94) | 0.99(0.93) | 1(0.97) | 1(0.96) | 0.99(0.96) |
| Slope of $X_1X_2(\beta_3)$ | 1(0.95) | 1(0.96) | 1(0.95) | 1(0.97) | 1.01(0.93) |
| Residual SD ($\sigma_\epsilon$) | 1.02 | 1.02 | 1 | 1.07 | 1.06 |
| $R^2$ | 0.74 | 0.74 | 0.75 | 0.71 | 0.72 |

Table 3.3: 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 $Y = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon$, where $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$ and $\epsilon \sim N(0, 1)$. The population coefficient of determination $R^2 = .75$.

Table 3.3 shows the estimates of $\alpha$, $\beta_1$, $\beta_2$, $\beta_3$, $\sigma_\epsilon$, $R^2$ and the coverage of the 95% confidence intervals for $\beta_1$, $\beta_2$ and $\beta_3$. 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 section 3.4.1. *'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 $\beta_1$, $\beta_2$ and $\beta_3$ 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 $\sigma_\epsilon$ and $R^2$ showed slight deviations from true values with MARtail and MARright.

### 3.4.3 Incomplete dataset with inequality restriction $\boldsymbol{X_1}+\boldsymbol{X_2} \geqq C$

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 $\epsilon$ 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 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.

### Results

| | 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 |

Table 3.4: 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 section 3.4.3. The true values of $E(X_1)$ and $E(X_2)$ are 0 and 3.5.

Table 3.4 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 [Morris et al., 2014].

## 3.5 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 $\sigma_\epsilon$ 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 section 3.4.

We limited our calculations and analyses to normal distributed $\boldsymbol{X}$. However, since Vink (2014) 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 $\boldsymbol{Y_{mis}}$ 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 section 3.2.3.

6. Repeat steps 1-5 $m$ times and save $m$ completed datasets.

# Part II

# Parametric imputation methods

# Chapter 4

# Identification of individual treatment effects by imputing potential outcomes with partial correlation specification

## Abstract

In most medical research, the treatment effect is calculated by the average treatment effect. However, precision medicine requires knowledge of individual treatment effects, which is how individuals respond to treatment. Because only one potential outcome is observed for each unit, the causal inference is a missing data problem. We proposed a multiple imputation approach to assess individual treatment effects, which allows the researcher to consider the heterogeneity of treatment effects unexplained by observed variables. The method needs to specify the partial correlation between potential outcomes since there is no relevant information in the observed data. Therefore, it is helpful to perform sensitivity analysis on the partial correlation. We showed that our approach provides valid inference of marginal distribution of potential outcomes. However, the posterior distribution of individual treatment effects varies with different specified partial correlations. According to the various posterior distribution of individual treatment effects, the researcher would comprehensively understand optimal treatment recommendations. We also apply our method to an HIV study and discuss further research directions.

## 4.1 Introduction

Heterogeneity of treatment effects (HTE) across individuals is an essential complication in precision treatment assignment to different persons. Attention to variability in treatment effects is increasing. If the true effect varies across individuals, the average treatment effect (ATE) leads to suboptimal individual treatment recommendations. The usual approach to assess HTE is subgroup analyses based on observed effect modifiers [Poulson et al., 2012, Zhang et al., 2013]. A subgroup analysis explains the HTE by observed data in different subpopulations. In fact, if all variables relevant to the heterogeneous treatment effect are available, we can build a model to estimate outcomes under all treatment levels. However, in practice, we rarely have all effect modifiers known and collected. In that case, there would be a fraction of HTE that cannot be explained by observed data, a situation known as idiosyncratic variation [Ding et al., 2016, Heckman et al., 1997] or residual HTE [Zhang et al., 2013].

In HTE, treatment effects vary by individual. The individual treatment effect (ITE) is the basic element in causal inference. We can always estimate average treatment effects, HTE or other statistics of interest from individual treatment effects. The potential outcomes framework stipulates that each individual has a potential outcome under each possible treatment level. Ding et al. [2018] reviewed various methods of estimating potential outcomes.

The difficulty of evaluating ITE from the observed data is that only one of the treatment outcomes is observed for each individual [Rubin, 1974, Hernan and Robins, 2010]. This fundamental problem of causal inference implies that causal inference is essentially a missing data problem [Rubin, 2005, Ding et al., 2018]. In this paper, we study the use of multiple imputation to quantify unobserved outcomes. In multiple imputation, each missing value is filled in multiple times to reflect the uncertainty about the true but unobserved value. After imputing missing values in potential outcomes, we can estimate the individual treatment effect on the completed datasets.

Lamont et al. [2018], and Westreich et al. [2015] discuss previous work on multiple imputation and potential outcomes. These authors fit the imputation model for a potential outcome based on observed pre-treatment variables, which we call multiple marginal imputation (MMI). MMI assumes conditional independence between potential outcomes. This assumption cannot be verified from the observed data [Rässler, 2012]. Imbens and Rubin [2015] and Gadbury et al. [2001] elucidate sensitivity of the estimates under violations of the assumption. van Buuren [2018] demonstrated that if one imputes the unobserved outcomes using only the observed outcome as a predictor, the derived imputed outcomes are unstable and implausible. Smink [2016] proposed a data augmentation approach to specify the correlation between potential outcomes. Smink focused on the situation without any covariates. With a proper prior specification of the joint distribution of potential outcomes, additional concatenated cases are generated and included in the observed data. In that case, the correlation between the imputed dataset's potential outcomes is close to the specified value and, hence, the imputed dataset is stable. However, it becomes challenging to specify the prior joint distribution for the augmented data if there are covariates. Proposed approaches for estimating potential outcomes are diverse. One of them is the targeted learning method, which is a machine learning based approach. It first fits the data-generating process among a set of reasonable models and then reduces the bias in estimating the parameter of interest iteratively. When evaluating the potential outcomes, the scientific interest is usually the average treatment effect.

This paper explores a new block imputation approach for missing potential outcomes with covariates. To clarify ideas, we focus on continuous outcomes. Potential outcomes are assumed with a multivariate normal distribution and imputed in one block. We introduce the notation, describe the potential outcomes framework and explore the role of the partial correlation between potential outcomes in causal inference in section 4.2. Section 4.3 provides an overview of the joint modeling method (JM) and the fully conditional specification method (FCS) and introduces block imputation, a hybrid of JM and FCS. We describe our proposed method, which imputes missing potential outcomes with partial correlation specification in section 4.4. To assess the validity of our method, we perform simulation studies and comparisons with the targeted learning in section 4.5. Section 4.6 demonstrates the applicability of our proposed approach in a clinical trial that aimed to compare the effect of four different therapies on slowing the progression of HIV disease. In section 4.7, we conclude with a discussion.

## 4.2 Framework

### 4.2.1 Notation

Let $Y(j), j = 1, \ldots, p$ denote one of $p$ incomplete variables and $Y(-j) = (Y(1), \ldots, Y(j-$ $1), Y(j+1),$

$\ldots, Y(p))$ denote the collection of the $p-1$ variables in $Y$ except $Y(j)$. In this paper, $Y(j)$ ususally represents the potential outcomes. Let $X = (X(1), \ldots, X(k))$ be a set of $k$ completely observed variables. Let $\rho(Y(0), Y(1))$ be the correlation between two potential outcomes and $\rho_{Y(0),Y(1)\mid\ X}$ be the partial correlation between two potential outcomes.

### 4.2.2 Setup

We focus on the case of a binary treatment $W_i$ and a continuous outcome $Y_i$ and assume the data come from a random sample of individuals, indexed by $i \in 1, \cdots, N$. Each individual $i$ has a nonzero probability to be assigned to both treatments, with $W_i = 1$ for the active treatment and $W_i = 0$ for the control treatment. The number of units under treatment and control are $N_1$ and $N_0$ respectively. We assume that the treatment assignment mechanism is unconfounded by the unobserved outcomes $Y_{mis}$, i.e., an ignorable assignment mechanism. We also assume that the potential outcomes for any individual are independent of the treatments assigned to others, which is known as the Stable Unit Treatment Value Assumption [Imbens and Rubin, 2015]. Here the ignorable or unconfounded assignment mechanism implies that $P(W|Y(0), Y(1), X) = P(W|Y_{obs}, X)$, where X are observed covariates not influenced by treatment assignment. The individual treatment effect is defined as $\tau_i = Y_i(1) - Y_i(0)$. We assume a joint distribution for the potential outcomes $Y_1$ and $Y_0$ and that the correlation between the potential outcomes $\rho(Y_1, Y_0)$ can be quantified by one or more parameters. The imputation models for missing outcomes

$Y(1)$ and $Y(0)$ are:

$$\dot{Y}(1) \sim P(Y^{mis}(1)|Y^{obs}(1), Y(0), X, \dot{\phi}_1) \tag{4.1}$$

$$\dot{Y}(0) \sim P(Y^{mis}(0)|Y^{obs}(0), Y(1), X, \dot{\phi}_0), \tag{4.2}$$

where parameters of the imputation model $\dot{\phi}_1$ and $\dot{\phi}_0$ are draws from their respective posterior distribution.

### 4.2.3 Partial correlation between potential outcomes

We decompose the individual treatment effect $\tau_i \in T, i = 1, \ldots, N$ via

$$\tau_i = Y_i(1) - Y_i(0) = X_i^T \beta + \epsilon_i, \tag{4.3}$$

where $X_i$ are observed pre-treatment covariates for individual $i$. Under random treatment assignment, the regression weight $\beta$ could be the ordinary least-square (OLS) estimation of $\tau_i$ on $X_i$. The quantity $X_i^T \beta$ is known as systematic treatment effect variation and the residual $\epsilon_i$ is the idiosyncratic treatment effect variation not explained by $X_i$ [Ding et al., 2019, Heckman et al., 1997, Djebbari and Smith, 2008]. The idiosyncratic variation accounts for treatment effect variation not attributable to differences in observed covariates. Based on the formula of OLS estimation, the coefficient

$$\begin{aligned}\beta &= (X^T X)^{-1} X^T T \\ &= (X^T X)^{-1} X^T Y(1) - (X^T X)^{-1} X^T Y(0) \\ &= \beta_1 - \beta_0,\end{aligned} \tag{4.4}$$

where $\beta_1$ and $\beta_0$ are the corresponding regression weights of the potential outcomes $Y_1$ and $Y_0$ on the observed covariates $X$. Similarly, the idiosyncratic treatment effect variation $\epsilon_i$

$$\begin{aligned}\epsilon_i &= \tau_i - X_i^T \beta \\ &= (Y_i(1) - X_i^T \beta_1) - (Y_i(0) - X_i^T \beta_0) \\ &= \epsilon_i(1) - \epsilon_i(0),\end{aligned} \tag{4.5}$$

where $\epsilon(1)$ and $\epsilon(0)$ are the residuals from the regression of the potential outcomes $Y_1$ and $Y_0$ on the observed covariates $X$. Applying the theory of variance decomposition for linear regression, we could decompose the variance of individual treatment effect into two components:

$$var(\tau_i) = var(X_i^T \beta) + var(\epsilon_i). \tag{4.6}$$

Based on the idiosyncratic treatment effect variation formula (5), the idiosyncratic components of individual treatment variance becomes

$$\begin{aligned} var(\epsilon_i) &= var(\epsilon_i(1) - \epsilon_i(0)) \\ &= var(\epsilon_i(1)) + var(\epsilon_i(0)) - 2cov(\epsilon_i(1), \epsilon_i(0)), \end{aligned} \tag{4.7}$$

which demonstrate that partial correlation between potential outcomes $\rho_{Y(0)Y(1)\mid X}$ does impact the idiosyncratic variation which is unidentifiable from the observed data.

We provide three approaches to identify the partial correlation between potential outcomes: the model-based approach, the experiment-based approach, and sensitivity analysis. The model-based approach explicitly models the relationship between the idiosyncratic variation and the assignment mechanisms such that the partial correlation would be close or equal to zero [Heckman, 2005]. Economists usually deal with ex-post causal inference. The agents select the treatment according to their ex-ante evaluation. In this case, economists could infer the information only available to agents from the assignment mechanism. For instance, Heckman et al. [2010] investigated the causal effect of educational decisions on the labor market and health outcomes. He modeled latent cognitive and social-emotional endowments and included these latent variables into the outcome equations. As indicated earlier, the partial correlation between potential outcomes is the only unknown parameter in the formula of idiosyncratic variation. The model for idiosyncratic variation could be tailored to the model for the partial correlation between potential outcomes. Generally, the model-based approach attempts to figure out latent variables affecting the partial correlation between potential outcomes.

The experiment-based approach intends to design sophisticated experiments to collect additional data where analysts could evaluate the partial correlation between potential outcomes. For instance, the experiment-based approach collects repeated measurements under more than one treatment level for the same individual from which some relevant information about the partial correlation between potential outcomes is available. The experiment designed for repeat measurements is known as N-1 trails [Shamseer et al., 2015, Araujo et al., 2016]. Researchers could also design an auxiliary treatment ($W_i = 2$) and assignment the extensive treatment to all individuals in the sample. Then the individual treatment effect $Y_i(1) - Y_i(0)$ can be evaluated by $(Y_i(2) - Y_i(0) - (Y_i(2) - Y_i(1))$.

Both model-based and experiment-based approaches search for extra information to determine the partial correlation between potential outcomes. If such additional information is not available, the alternative is sensitivity analysis. One could impute the missing potential outcomes and evaluate individual treatment effects with various valid partial correlations and study the effect on the conclusions [Gadbury et al., 2001].

## 4.3 Methodology

### 4.3.1 Joint modeling imputation (JM)

Joint modeling imputation assumes a model $p(Y^{mis}, Y^{obs} \mid \theta)$ for the complete data and a prior distribution $p(\theta)$ for the parameter $\theta$. Joint modeling partitions the observed data into groups based on the missing pattern and imputes the missing data within each missing pattern according to corresponding predictive distribution. Under the assumption of ignorability, the parameters of the predictive distribution for different missing patterns are generated from the posterior joint distribution. Schafer [1997] proposed joint modeling methods for multivariate normal data, categorical data and mixed normal-categorical data. The joint modeling approach has solid theoretical properties (i.e., compatibility between the imputation and substantive models) while it lacks the flexibility of model specification.

### 4.3.2 Fully Conditional Specification (FCS)

In Fully Conditional Specification, we specify the distribution for each partially observed variable conditional on all other variables $P(Y(j)|Y(-j), X, \theta_j)$ and impute each missing variable iteratively. The FCS starts with naive imputations such as a random draw from the observed values. The $t$th iteration for the incomplete variable $Y(j)$ consists of the following draws:

$$\theta_j^t \sim f(\theta_j) f(Y^{obs}(j)|Y^{t-1}(-j), X, \theta_j)$$
$$Y^{mis(t)}(j) \sim f(Y^{mis}(j)|Y^t(-j), X, \theta_j^t),$$

where $f(\theta_j)$ is generally specified as a noninformative prior. After a sufficient number of iteration, typically with 5 to 10 iterations [van Buuren, 2018, Oberman et al., 2020], the stationary distribution is achieved. The final iteration generates a single imputed dataset and the multiple imputations are created by applying FCS in parallel $m$ times. Since FCS provides tremendous flexibility in specifying imputation models for multivariate partially observed data, FCS is now a widely accepted and popular MI approach [Van Buuren, 2007]. Even while, FCS lacks a satisfactory theory and has a potential risk in incompatibility.

### 4.3.3 Block imputation

Block imputation combines the flexibility of FCS with the attractive theoretical properties of JM. A block consists of one or more variables. If the block has multiple variables, then we use multivariate imputation methods to impute those variables jointly. A simple example would be multiple imputation of missing variables with quadratic effects $Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$. In such a case, grouping the missing variable $X$ and its corresponding square term $X^2$ within one block is of benefit to preserve the quadratic relationship [Vink and van Buuren, 2019]. The joint modeling approach is the special case where all variables form one block, while the FCS approach treats each variable as a separate block.

When the imputation model of one variable is potentially incompatible, or its theoretical properties are not fully studied (i.e., whether the imputations based on the FCS correspond to drawing from a joint distribution), block imputation would merge that variable with other variables and apply the joint modeling imputation approach to that block. On the other hand, when the joint distribution of several missing variables is ambiguous, block imputation could use the FCS approach to impute each variable. In general, the apparent advantage of block imputation is the flexibility of model specification. However, block methods are hardly known or studied. While available in the `mice` software, the properties of block imputation have yet to be studied.

## 4.4 BLI-SPC

This section describes how the block imputation could be used to impute missing outcomes with a given partial correlation between potential outcomes. We term the algorithm the block imputation algorithm with specified partial correlation (BLI-SPC). Since the missing pattern of potential outcomes is somehow restrictive, that is, no cases with completely observed potential outcomes, the imputation procedure is comprised of three steps. The first step involves estimating the marginal distribution of potential outcomes conditioning on pre-treatment variables. The second step consists of deriving the multivariate density of potential outcomes by combining the marginal distribution of potential outcomes and the specified correlation between potential outcomes. The third step imputes the missing outcomes with the corresponding submodel obtained from the multivariate distribution.

Our approach shares some similarities with statistical matching discussed by Moriarity and Scheuren [2003]. Suppose there are two sample files, A and B. File A collects variables $X$ and $Y$ and file B collects variables $X$ and $Z$. The purpose of statistical matching is to combine two files, A and B, into one file containing variables $X$, $Y$ and $Z$. Rubin [1986] proposed a procedure of statistical matching with three steps: regression step, matching step and concatenation step. In the regression step, Rubin specified the correlation between variable $X$ and $Y$ to derive the joint distribution of ($X$, $Y$, $Z$) in two sample files. We develop this idea to evaluate the individual treatment effects and extend to multiple treatments condition.

Without loss of generality, let us assume that potential outcomes follow a multivariate normal distribution. We specify Bayesian linear models for two potential outcomes based on observed covariates.

$$Y(0) = \beta_0 X + \epsilon_0, \epsilon_0 \sim \mathcal{N}(0, \sigma_0^2) \tag{4.8}$$

$$Y(1) = \beta_1 X + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \tag{4.9}$$

Bayesian sampling draws $\beta_0^*$, $\beta_1^*$, $\sigma_0^{*2}$, $\sigma_1^{*2}$ from their respective posterior distribution. The Jeffrey's prior used and hence, the posterior distributions of $\sigma_0^{*2}$ and $\sigma_1^{*2}$ would be inverse $\chi^2$ distribution:

$$\sigma_0^{*2} \sim \sum_{i=1}^{N_0} (Y_i(0) - \hat{\beta}_0 X_i)^2 \chi_{N_0-k}^{-2} \tag{4.10}$$

$$\sigma_1^{*2} \sim \sum_{i=1}^{N_1} (Y_i(1) - \hat{\beta}_1 X_i)^2 \chi_{N_1-k}^{-2} \tag{4.11}$$

where $\hat{\beta}_0 = (X_0'X_0)^{-1}X_0'Y(0)$, $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y(1)$ and $k$ is the number of covariates. The conditional distributions of $\beta_0^*$ and $\beta_1^*$ are multivariate normal:

$$\beta_0^* | \sigma_0^{*2} \sim \mathcal{N}(\hat{\beta}_0, \sigma_0^{*2}((X_0'X_0)^{-1})) \tag{4.12}$$

$$\beta_1^* | \sigma_1^{*2} \sim \mathcal{N}(\hat{\beta}_1, \sigma_1^{*2}((X_1'X_1)^{-1})) \tag{4.13}$$

Since there is no information relevant to the partial correlation between potential outcomes in the observed data, the posterior distribution of the partial correlation $\rho_{Y(0)Y(1)|X}$ equals the prior distribution specified by the user, who can select several numbers in the interval $[-1, 1]$ to investigate the sensitivity to $\rho_{Y(0)Y(1)|X}$. Finally, combined the marginal distribution of $Y(0)$ and $Y(1)$ with the specification of $\rho_{Y(0)Y(1)|X}$, the joint distribution of $Y^{mis}(0)$ and $Y^{mis}(1)$ is:

$$\begin{bmatrix} Y^{mis}(0) \\ Y^{mis}(1) \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \beta_0^* x \\ \beta_1^* x \end{pmatrix}, \begin{pmatrix} \sigma_0^{*2} & \rho_{Y(0)Y(1)|X}\sigma_0^*\sigma_1^* \\ \rho_{Y(0)Y(1)|X}\sigma_0^*\sigma_1^* & \sigma_1^{*2} \end{pmatrix}\right] \tag{4.14}$$

and the distributions of $Y^{mis}(0)$ and $Y^{mis}(1)$ are:

$$Y^{mis}(0) \sim \mathcal{N}(\beta_0^* x + (Y(1) - \beta_1^* x)\rho_{Y(0)Y(1)|X}\sigma_0/\sigma_1, (1 - \rho^2_{Y(0)Y(1)|X})\sigma_0^{*2}) \tag{4.15}$$

$$Y^{mis}(1) \sim \mathcal{N}(\beta_1^* x + (Y(0) - \beta_0^* x)\rho_{Y(0)Y(1)|X}\sigma_1/\sigma_0, (1 - \rho^2_{Y(0)Y(1)|X})\sigma_1^{*2}) \tag{4.16}$$

Comparing equations (15) and (16) to equations (8) and (9), it is evident that inclusion of the observed outcome may change the location of missing outcomes shifts slightly and the uncertainty is reduced when imputing missing outcomes under the specified correlation between potential outcomes. For the prediction of units out of trials, the reasonable values for outcomes under two treatments could be drawn from the joint distribution (14).

When generalizing to the multiple treatments condition $W = 0, 1, \ldots, w$, the marginal posterior distribution for potential outcomes would be:

$$\begin{gathered} Y(0)^* = \beta_0^* X + \epsilon_0^*, \epsilon_0^* \sim \mathcal{N}(0, \sigma_0^{*2}) \\ Y(1)^* = \beta_1^* X + \epsilon_1^*, \epsilon_1^* \sim \mathcal{N}(0, \sigma_1^{*2}) \\ \cdots \\ Y(w)^* = \beta_w^* X + \epsilon_w^*, \epsilon_w^* \sim \mathcal{N}(0, \sigma_w^{*2}) \end{gathered}$$

where the values of $\beta_0^*$, $\beta_1^*$, ..., $\beta_w^*$, $\sigma_0^{*2}$, $\sigma_1^{*2}$, ..., $\sigma_w^{*2}$ draw from their respective Bayesian posterior distribution. If $\sigma_0^{*2}, \ldots, \sigma_w^{*2}$ are unrestricted, with pairwise specification of partial correlation between potential outcomes, the joint distribution of $Y^{mis}(0)$, $Y^{mis}(1)$,

…, $Y^{mis}(w)$ is:

$$\begin{bmatrix} Y^{mis}(0) \\ Y^{mis}(1) \\ \cdots \\ Y^{mis}(w) \end{bmatrix} \sim \mathcal{N}\Big[\ \mathcal{M}\ ,\ \Sigma\ \Big], \tag{4.17}$$

where $\mathcal{M} = (\beta_0^* x, \beta_1^* x, \ldots, \beta_w^* x)^T$. The covariance matrix $\Sigma$ must be positive semi-definite:

$$\Sigma \ = \begin{pmatrix} \sigma_0^{*2} & \rho_{Y(0)Y(1)|X}\sigma_0^*\sigma_1^* & \cdots & \rho_{Y(0)Y(w)|X}\sigma_0^*\sigma_w^* \\ \rho_{Y(1)Y(0)|X}\sigma_1^*\sigma_0^* & \sigma_1^{*2} & \cdots & \rho_{Y(1)Y(w)|X}\sigma_1^*\sigma_w^* \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{Y(w)Y(0)|X}\sigma_w^*\sigma_0^* & \rho_{Y(w)Y(1)|X}\sigma_w^*\sigma_1^* & \cdots & \sigma_w^{*2} \end{pmatrix} \tag{4.18}$$

Draws of missing outcomes for units under different treatments could be derived from the joint distribution based on the property of conditional distribution for the multivariate normal distribution. For instance, with units under control treatment $W = 0$, the distribution of missing outcomes $Y^{mis}(-0) = (Y^{mis}(1), \ldots, Y^{mis}(w))$ would be $Y^{mis}(-0) \sim \mathcal{N}((\beta_1^* x, \ldots, \beta_w^* x)^T + \Sigma_{0-0}\Sigma_{-0-0}^{-1}(Y_0 - \beta_0^* x), \Sigma_{00} - \Sigma_{0-0}\Sigma_{-0-0}^{-1}\Sigma_{-00})$, where

$$\begin{bmatrix} \Sigma_{00} & \Sigma_{0-0} \\ \Sigma_{-00} & \Sigma_{-0-0} \end{bmatrix} \tag{4.19}$$

is the partition of $\Sigma$: $\Sigma_{00} = \sigma_0^{*2}$, $\Sigma_{0-0} = \Sigma_{-00}^T = (\rho_{Y(0)Y(1)|X}\sigma_0^*\sigma_1^*, \ldots, \rho_{Y(0)Y(w)|X}\sigma_0^*\sigma_w^*)$ and

$$\Sigma_{-0-0} \ = \begin{pmatrix} \sigma_1^{*2} & \cdots & \rho_{Y(1)Y(w)|X}\sigma_1^*\sigma_w^* & \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{Y(w)Y(1)|X}\sigma_w^*\sigma_1^* & \cdots & \sigma_w^{*2} & \end{pmatrix} \tag{4.20}$$

One could use the sweep operator for rapid calculation of the parameters for imputation models of missing outcomes [Goodnight, 1979].

## 4.5 Simulation study

We evaluate the performance of BLI-SPC at both the individual level and the aggregate level. Specifically, we analyze the average bias of individual treatment effects and imputations for randomly selected 50 units from one repetition and investigate the quality of the estimated parameters for the distribution of the potential outcomes. We perform a sensitivity analysis to the multiple imputation approach with three different values of partial correlation between potential outcomes: $\rho = 0, 0.73$ and 0.99, which corresponding to, respectively, a conditional independent correlation assumption, the correct partial correlation and a constant treatment effect condition. Van der Laan and Rose [2011] developed the target learning to estimate the average treatment effect. We select the target learning method as a comparative approach to identify individual treatment effects because the target learning involves calculating the missing potential outcomes. However, unlike our proposed method, the target learning fits the distribution of the missing outcome only based on the covariates. In the simulation study, we aim to show it is necessary to consider the correlation between potential outcomes when specifying the analysis model of potential outcomes. This section will introduce the design of simulation, provide an overview of the targeted learning method and present the simulation results.

### 4.5.1 Simulation conditions

We design the two potential outcomes $Y(0)$ and $Y(1)$ as well as one baseline covariate $X$. The data is generated with a multivariate normal distribution:

$$\begin{pmatrix} Y_0 \\ Y_1 \\ X \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \right]$$

Because the marginal correlation between potential outcomes is 0.8, the corresponding partial correlation is 0.73.

$$\begin{aligned}\rho_{y_0 y_1|x} &= \tfrac{\rho_{y_0 y_1} - \rho_{y_0 x}\rho_{y_1 x}}{\sqrt{1-\rho^2_{y_0 x}}\sqrt{1-\rho^2_{y_1 x}}} \\ &= \tfrac{0.8-0.25}{0.75} \\ &\approx 0.73,\end{aligned} \tag{4.21}$$

A total of $N = 5000$ independent and identically distributed cases are generated. The first 2500 cases have only information for $Y_0$ and the remaining cases have only information on $Y_1$. One thousand repetitions of the simulation are produced. For reasons of brevity, we only include one pre-treatment covariate. However, it is straightforward to extend the methodology to situations with more covariates and even mixtures of continuous and categorical predictors.

### 4.5.2 Targeted learning

Targeted learning is an alternative for estimating individual treatment effects. The idea is to estimate the data-generating distribution $P_0$ and then update the initial estimation to make an optimal bias-variance tradeoff for the scientific interest $\Psi(P_0)$. To estimates individual treatment effects, we define the average treatment effect as the scientific interest:

$$\Psi(P_0) = E[E(Y_i(1) \mid X) - E(Y_i(0) \mid X)]. \tag{4.22}$$

The targeted learning consists of three steps of analysis: 1) definition of the data-generating model and the scientific interest $\Psi(P_0)$, 2) super learning for initial prediction of $\Psi(P_0)$ and 3) targeted maximum likelihood estimation for $\Psi(P_0)$ [Van der Laan and Rose, 2011]. Specifically, before estimation, it is necessary to define a set of possible probability distributions of observed data and identify a collection of causal assumptions (i.e., the ignorable assignment mechanism and the stable unit treatment value assumption) to the identification of the correct model. With the definition of the model, one could apply the super learner to derive an initial estimation for the distribution of potential outcomes $\hat{P}_0$. The super learner first selects a library of candidate algorithms and a risk function and then applies the validation set approach to calculate the average risk for each algorithm. The optimal algorithm is chosen with the smallest average risk and used to produce the initial predicted distribution of potential outcomes. The candidate algorithms could be parametric(i.e., general linear model), non-parametric (i.e., random forest), or even a weighted combination of statistical algorithms.

After the initial estimation of predicted potential outcomes, one could define the targeted maximum likelihood estimation (TMLE) for scientific interest $\Psi(P_0)$. The TMLE step reduces the bias in the estimation of $\Psi(P_0)$ if the initial estimation $\Psi(\hat{P}_0)$ is inconsistent. This is accomplished by exploiting information in the treatment assignment mechanisms to adjust the initial estimations. Generally, the adjustment is an iterative procedure. However, when the scientific interest is the average treatment effect, the convergence is achieved in one step. More details are provided in Gruber and Van der Laan [2011].

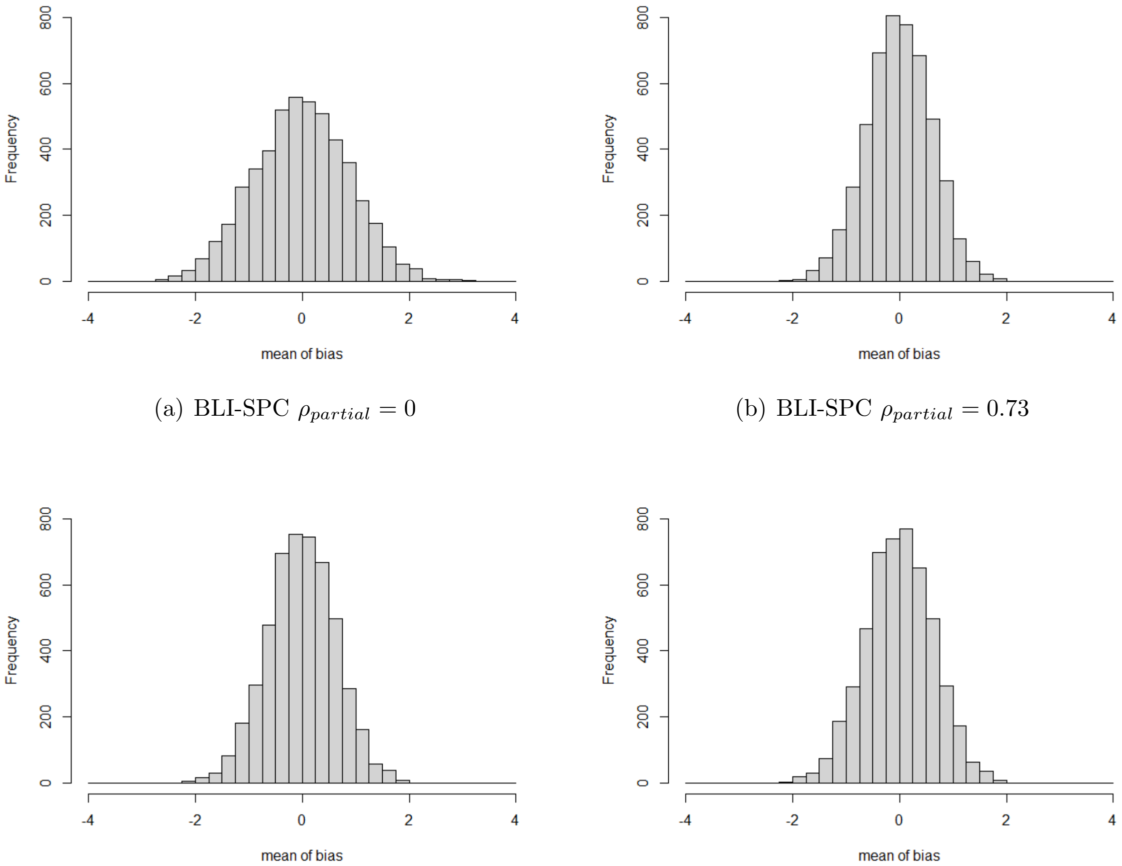
### 4.5.3 Results

**Predicted individual treatment effect**

We investigate the predictive accuracy of the individual treatment effect among multiple imputation with different partial correlations and targeted learning. Twenty repetitions of the imputed data are created. The true data is generated according to the multivariate normal distribution introduced in section 4.5.1 and set to be the same across repetitions. We evaluate the bias as the mean difference between true and predicted values and the variance of imputed outcomes for each individual across all repetitions.

Figure 4.1 shows the distribution of the mean bias for all four strategies, and the corresponding location and scale are displayed in Table 4.1.

Overall, BLI-SPC with three different partial correlations and the targeted learning all yield unbiased estimates of the average treatment effect. However, in terms of the scale, The BLI-SPC with the “correct” partial correlation has the smallest variance. The closer the specified partial correlation is to the “true” value, the smaller bias and



(a) BLI-SPC $\rho_{partial} = 0$

(b) BLI-SPC $\rho_{partial} = 0.73$

(c) BLI-SPC $\rho_{partial} = 0.99$

(d) The targeted learning

Figure 4.1: Histrogram plots of the mean bias for all four strategies to estimate the individual treatment effects.

| $\rho_{Y(0)Y(1)\mid X}$ | Mean | Variance |
|---|---|---|
| BLI-SPC $\rho_{partial} = 0$ | -0.012 | 0.791 |
| BLI-SPC $\rho_{partial} = 0.73$ | -0.007 | 0.363 |
| BLI-SPC $\rho_{partial} = 0.99$ | -0.013 | 0.398 |
| The targeted learning | -0.006 | 0.401 |

Table 4.1: The location and the scale of the mean bias for all four strategies to estimate the individual treatment effects.

variance would be produced. When the partial correlation is specified accurately, the variance of the bias of the individual treatment effect across all cases equals the partial variance of the potential outcome. It is thus challenging to produce an accurate estimate of the individual treatment effect for a unit at the extreme. We can see this in Figure 4.2. However, we still have a large proportion of the estimated individual treatment effect with negligible biases. Since the variance of the bias equals the partial variance

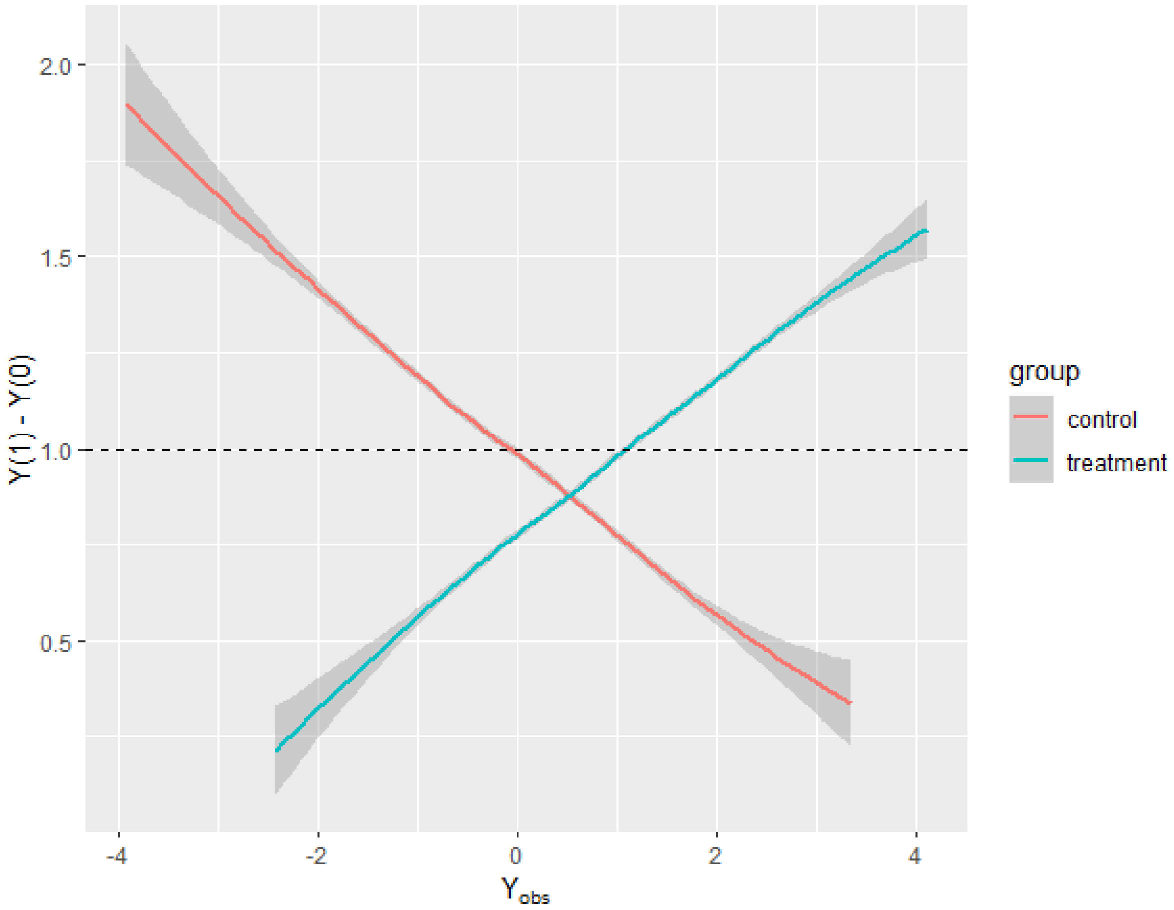
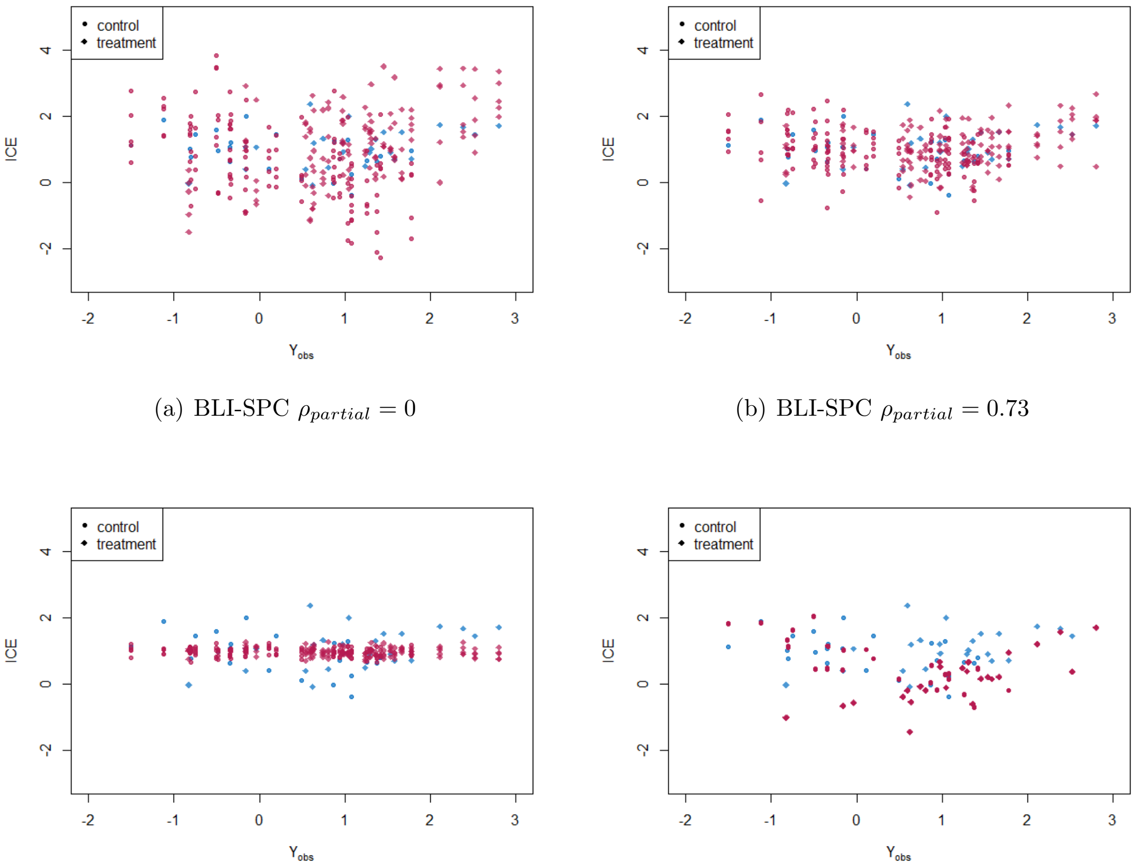


Figure 4.2: The plot shows individual treatment effects $(Y(1) - Y(0))$ in both treatment and control group. The oucome $Y(0)$ is observed in the control group and the oucome $Y(1)$ is observed in the treatment group. The dashed line represents the average treatment effect.

of the potential outcomes, we could include more explanatory variables to increase the accuracy of the imputation of missing outcomes and hence the prediction of individual treatment effect. On the other hand, when the specified partial correlation deviates from the "true" value, more variance would appear because of the differences between the true distribution and the estimated distribution for each missing outcome.

Figure 4.3 shows individual treatment effects calculated with imputed datasets for selected cases($i = 100, 200, \ldots, 5000$). When the partial correlation is specified correctly, the imputations look plausible: imputed ICE covers the true ICE for almost every case, and the variance of ICE for each individual is smaller than the case under the independent conditional correlation assumption. With homogeneous treatment effect assumption, i.e., $\rho_{Y(0)Y(1)\mid X} = 0.99$, the imputed individual treatment effects are biased towards the



(a) BLI-SPC $\rho_{partial} = 0$

(b) BLI-SPC $\rho_{partial} = 0.73$

(c) BLI-SPC $\rho_{partial} = 0.99$

(d) The targeted learning

Figure 4.3: Stripplot of $\mathtt{m} = 5$ of observed (blue) and imputed (red) data for individual treatment effects with selected cases.

average treatment effect. For targeted learning, uncertainty about the missing outcomes is not estimated. The BLI-SPC approach derives the distribution of individual treatment effects, which provides more information on treatment recommendations. For instance, with a small individual treatment effect, it is possible to estimate the probability of a positive causal effect from the distribution of individual treatment effect.

### Parameter inference

In this section, we investigate whether we could provide valid inferences of the distribution of the potential outcomes. The data generation mechanism follows the multivariate normal distribution introduced in section 4.5.1. The simulation is repeated 1000 times. We are interested in the biases and the coverage of nominal 95% CIs of all parameters connected to the potential outcomes. Table 4.2 shows all statistics relevant to the potential outcomes. Statistics involving only one potential outcome are unbiased and have valid coverage rates, which means that even with incorrect specified partial correlation, we could derive plausible marginal distribution of potential outcomes. Since the partial correlation is set before imputation and there is no information about the correlation between potential outcomes in the data, we get valid inference for the marginal correlation between potential outcomes only when we specified the partial correlation correctly.

| Method | Truth | Est | Cover |
|---|---|---|---|
| BLI-SPC $\rho_{partial} = 0$ | | | |
| E($y_0$) | 0.0 | 0.00 | 0.95 |
| E($y_1$) | 1.0 | 1.00 | 0.94 |
| Var($y_0$) | 1.0 | 1.00 | 0.94 |
| Var($y_1$) | 1.0 | 1.00 | 0.94 |
| Cov($y_0$, $y_1$) | 0.8 | 0.25 | 0.00 |
| Cov($y_0$, $x$) | 0.5 | 0.50 | 0.94 |
| Cov($y_1$, $x$) | 0.5 | 0.50 | 0.95 |
| BLI-SPC $\rho_{partial} = 0.73$ | | | |
| E($y_0$) | 0.0 | 0.00 | 0.97 |
| E($y_1$) | 1.0 | 1.00 | 0.95 |
| Var($y_0$) | 1.0 | 1.00 | 0.94 |
| Var($y_1$) | 1.0 | 1.00 | 0.95 |
| Cov($y_0$, $y_1$) | 0.8 | 0.80 | 1.00 |
| Cov($y_0$, $x$) | 0.5 | 0.50 | 0.95 |
| Cov($y_1$, $x$) | 0.5 | 0.50 | 0.94 |
| BLI-SPC $\rho_{partial} = 0.99$ | | | |
| E($y_0$) | 0.0 | 0.00 | 0.95 |
| E($y_1$) | 1.0 | 1.00 | 0.95 |
| Var($y_0$) | 1.0 | 1.00 | 0.95 |
| Var($y_1$) | 1.0 | 1.00 | 0.94 |
| Cov($y_0$, $y_1$) | 0.8 | 0.99 | 0.00 |
| Cov($y_0$, $x$) | 0.5 | 0.50 | 0.95 |
| Cov($y_1$, $x$) | 0.5 | 0.50 | 0.95 |

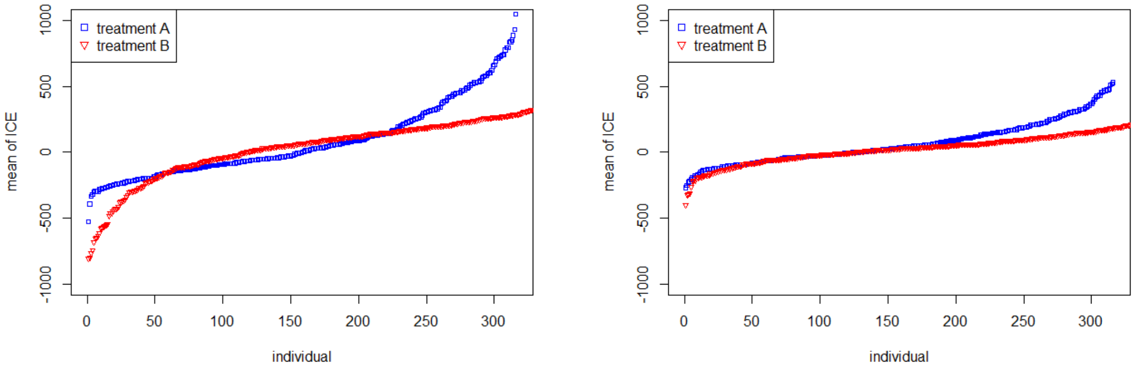
Table 4.2: Parameter estimates for BLI-SPC with different partial correlations

## 4.6 Application

We applied our approach to evaluate the effects of two different therapies on slowing the progression of HIV disease. The data came from a study comparing the effects of four therapies (zidovudine alone, didanosine alone, zidovudine plus didanosine and zidovudine and zalcitabine) on preventing the deterioration of disease in adults with HIV-1 infected patients [Hammer et al., 1996]. For simplicity, we name them treatment A(zidovudine alone), B(didanosine alone), C(zidovudine plus didanosine) and D(zidovudine and zalcitabine). The data named ACTG175 is accessible in package speff2trial in R. We restrict our analyses to treatments A and B and perform out-of-sample prediction of hypothetical effect of treatments A and B for the remaining patients that were allocated to treatments C and D. Hammer et al. [1996] concluded that treatment with didanosine is superior to treatment with zidovudine. However, the overall treatment effect was found to be insufficient to recommend the therapy to a patient, a situation that in common in many medical interventions.

A total of 693 HIV-1 infected adults with CD4 cell counts in the range of 200 to 500 per cubic millimeter were randomized into control (N = 316) and treatment (N = 377) group, while 670 patients are treated as out-of-sample. Fifteen baseline covariates are included which assess gender, age, weight, Karnofsky score, risk factors, prior antiretroviral therapy, CD4 cell count, and CD8 T cell count. We are interested in the number of days until the first occurrence of: 1) a decline in CD4 cell count of at least 50 2) an event indicating progression to AIDS, or 3) death. The larger number of days yields a more beneficial treatment effect. The individual treatment effect is defined as the number of days under treatment B minus the number of days under treatment A. We select the value of partial correlation as 0 and 0.7 to perform the sensitivity analysis so that the result yields distinct differences.

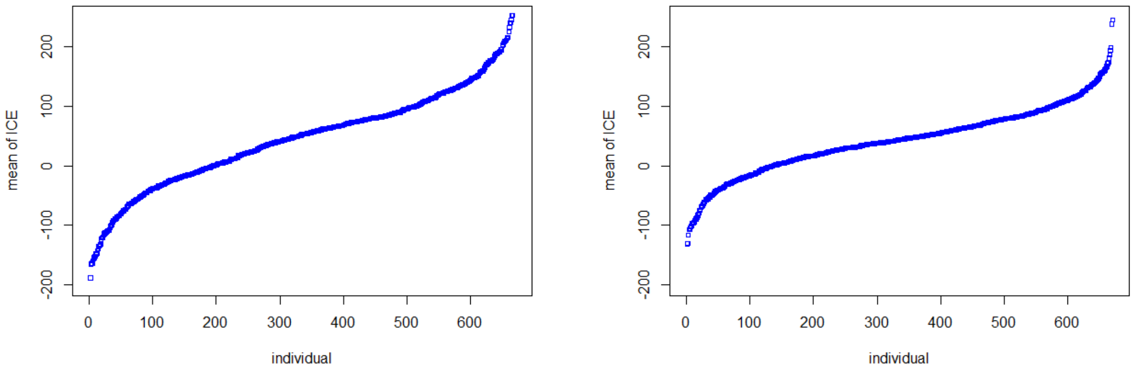
Figure 4.4 shows the results of individual treatment effects under treatment A and treatment B group with partial correlation specification 0 and 0.7. As expected, this approach could detect the heterogeneity of treatment effects. A large proportion of patients in the sample, whose individual treatment effects are larger than 0, are recommended to receive a treatment regimen with didanosine. However, treatment with zidovudine still yields greater clinical benefit for a fraction of units, whose individual treatment effects are smaller than 0. Since all covariates are balanced under two groups, the distributions of expected individual treatment effects under two treatments (A and B) are more similar



(a) BLI-SPC $\rho_{partial} = 0$

(b) BLI-SPC $\rho_{partial} = 0.7$

Figure 4.4: Means of individual treatment effects for patients under treatment A and treatment B, which are arranged into ascending order.



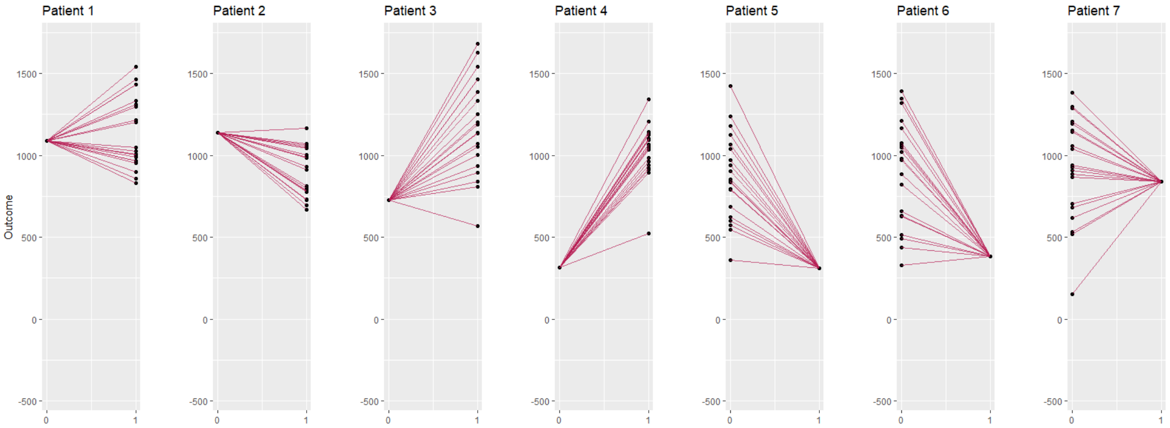
(a) BLI-SPC $\rho_{partial} = 0$

(b) BLI-SPC $\rho_{partial} = 0.7$

Figure 4.5: Means of individual treatment effects for out-of-sample patients under treatment C and D, which are arranged into ascending order.

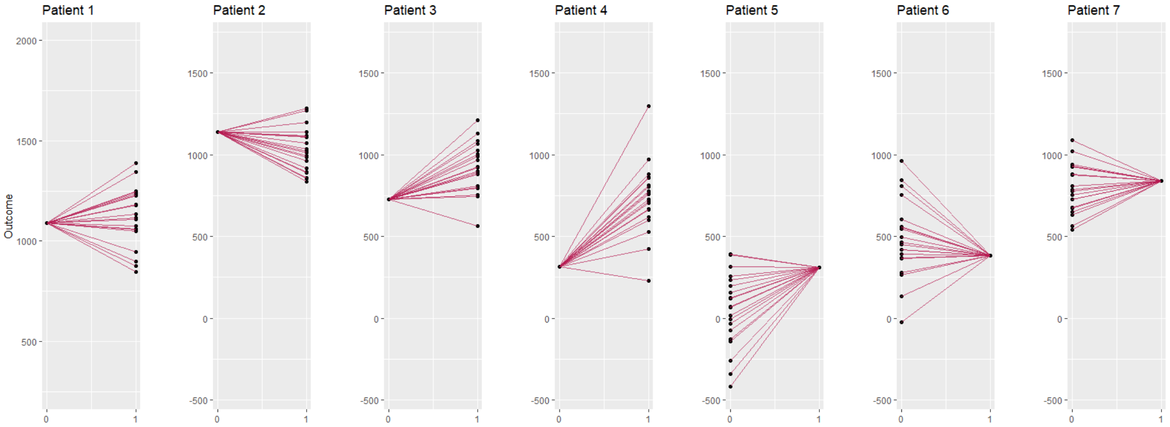
when specifying a 0.7 partial correlation. Furthermore, the range of expected value of individual treatment effects is smaller, with a partial correlation of 0.7. The larger the partial correlation we set, the more convinced that all effect modifiers are included, and effects are identical across persons. Figure 4.5 shows variability in individual effects in out-of-sample patients, which implies that our method could also be applied to prediction.

Figure 4.6 and 4.7 display imputations by selected patients for two different values of partial correlation, 0 and 0.7. Each panel contains the observed outcome for the patient



(a) (b) (c) (d) (e) (f) (g)

Figure 4.6: Fan plot of Observed and imputed (m = 20) outcomes under treatment A (0) and B (0). The partial correlation is 0.



(a) (b) (c) (d) (e) (f) (g)

Figure 4.7: Fan plot of Observed and imputed (m = 20) outcomes under treatment A (0) and B (0). The partial correlation is 0.7.

and $m = 20$ imputed values for the missing outcome. The imputed outcomes are sensitive to different values of partial correlation. Patient 5 benefits from treatment A when the partial correlation equals 0, while we derive the opposite conclusion when specifying the partial correlation as 0.7. The location of imputed outcomes is different under various partial correlations, and the scale shrinks when the partial correlation tends to 1.

## 4.7 Discussion

Treatment assignment is currently steered by the average treatment effect. This may lead to the suboptimal individual treatment decision when the assumption of homogeneous treatment effect is not true. More often than not, we lack the explanatory variables that would render the ATE homogeneous. This paper proposes a multiple imputation approach to replace missing outcomes with plausible values to estimate the individual treatment effect. Our method allows researchers to consider the correlation between potential outcomes, which influence the imputation of missing outcomes.

The sensitivity analysis of the partial correlation between potential outcomes in section 4.5.3 demonstrates that different values of the partial correlation yielded similar results in terms of marginal distributions of potential outcomes and the average treatment effect. However, the closer the specified partial correlation is to the "true" value, the less biased the estimated individual treatment effect will be. Since one cannot obtain information about the partial correlation from the observed data, the determination of the partial correlation should be set to a plausible range based on previous investigations or expert knowledge. In addition, it may be useful to perform a sensitivity analysis to see how imputed outcomes differ with different partial correlations.

An advantage of our multiple imputation approach to individual treatment effects is that it provides an estimate of the uncertainty of the imputed outcomes and hence, of the individual treatment effects. One could obtain the posterior distribution of the individual treatment effects for a unit from multiply imputed datasets, from which we can learn the probability of benefit from the treatment at the individual level. Since we incorporate the partial correlation when imputing, researchers could apply complete-data analyses to explore potential variables. This is accounting for residual heterogeneity of treatment effects or additional effect modifiers.

In our illustration of the BLI-SPC algorithm, we applied Bayesian imputation under the normal linear model. It is possible to use other imputation techniques (parametric or non-parametric imputation methods). The behavior of such methods have not yet been studied. Another useful property of BLI-SPC is that under the assumption of ignorable treatment assignment, researchers can skip explicit modeling of the probability of assignment. The BLI-SPC is a hybrid of FCS and JM, and hence it also provides valid imputations with MAR.

In the application study, we focus on the comparison of treatments A and B. It is possible to generalize to the multiple treatment comparison (treatment A, B, C, and D). By imputing unobserved outcomes, we could recommend the optimal treatment to each unit among four treatments. One could benefit from our method when performing an experiment that has been investigated on a different population. Some proven effect modifiers may be difficult to collect when performing the same experiment in other regions or countries. In such a case, the inference would include the heterogeneity of treatment effects explained by the uncollected factors by specifying a reasonable partial correlation.

This is an initial study on MI to the individual treatment effect. The simulation study used a basic randomized trial with a correctly specified imputation model. Further work should be done to extend discrete and semi-continuous outcomes. Another challenge is to develop imputation techniques for studies that collect post-treatment variables. All in all, we believe that our methodology for incorporating the partial correlation represents an important advance in estimating individual treatment effects. We hope that our method may attribute to put down the shackles of the ubiquitous average treatment effect.

# Chapter 5

# Joint distribution properties of Fully Conditional Specification under the normal linear model with normal inverse-gamma priors

## Abstract

Fully conditional specification (FCS) is a convenient and flexible multiple imputation approach. It specifies a sequence of simple regression models instead of a potential complex joint density for missing variables. However, FCS may not converge to a stationary distribution. Many authors have studied convergence properties of FCS when priors of conditional models are non-informative. We extend to the case of informative priors. This paper evaluates the convergence properties of the normal linear model with normal-inverse gamma prior. The theoretical and simulation results prove the convergence of FCS and show the equivalence of prior specification under the joint model and a set of conditional models when the analysis model is a linear regression with normal inverse-gamma priors.

## 5.1 Introduction

Multiple imputation [Rubin, 1987] is a widely applied approach for the analysis of incomplete datasets. It involves replacing each missing cell with several plausible imputed values that are drawn from the corresponding posterior predictive distributions. There are two dominant approaches to arrive at those posterior distributions under multivariate missing data: joint modeling (JM) and fully conditional specification (FCS).

Joint modeling requires a specified joint model for the complete data. Schafer [1997] illustrated joint modeling imputation under the multivariate normal model, the saturated multinomial model, the log-linear model, and the general location model. However, with an increasing number of variables and different levels of measurements, it can be challenging to formulate the joint distribution of the data.

Fully conditional specification offers a solution to this challenge by allowing a flexible specification of the imputation model for each partially observed variable. The imputation procedure then starts by imputing missing values with a random draw from the marginal distribution. Each incomplete variable is then iteratively imputed with a specified univariate imputation model.

Fully conditional specification has been proposed under a variety of names: chained equations stochastic relaxation, variable-by-variable imputation, switching regression, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC and iterated univariate imputation [van Buuren, 2018, Section 4.5.1]. Fully conditional specification can be of great value in practice because of its flexibility in model specification. FCS has become a standard in practice and has been widely implemented in software (e.g. `mice` and `mi` in `R`, `IVEWARE` in `SAS`, `ice` in `STATA` and module `MVA` in `SPSS`) [van Buuren and Groothuis-Oudshoorn, 2011].

Although many simulation studies demonstrated that fully conditional specification yields plausible imputations in various cases, the theoretical properties of fully conditional specification are not thoroughly understood [Van Buuren, 2007]. A sequence of conditional models may not imply a joint distribution to which the algorithm converges. In such a case, the imputation results may systematically differ according to different visit sequences, which is named as “order effects” [Hughes et al., 2014].

van Buuren [2018, Section 4.6.1] stated two cases in which FCS converges to a joint distribution. First, if all imputation models are linear with a homogenous normal distributed response, the implicit joint model would be the multivariate normal distribution. Second, if three incomplete binary variables are imputed with a two-way interactions logistic regression model, FCS would be equivalent to the joint modeling under a zero three-way interaction log-linear model. Liu et al. [2014] illustrated a series of sufficient conditions under which the imputation distribution for FCS converges in total variation to the posterior distribution of a joint Bayesian model when the sample size moves to infinity. Complementing the work of Liu et al. [2014], Hughes et al. [2014] pointed out that, in addition to the compatibility, a “non-informative margins” condition is another sufficient condition for the equivalency of FCS and joint modeling for finite samples. Hughes et al. [2014] also showed that with multivariate normal distributed data and a non-informative prior, both compatibility and the non-informative margins conditions are satisfied. In that case, fully conditional specification and joint modeling provide imputations from the same posterior distribution. Zhu and Raghunathan [2015] discussed conditions for convergence and assess properties of FCS. Many authors illustrated convergence properties of FCS when the prior for conditional models is non-informative. However, the case of informative priors has not received much attention. Therefore, we should consider the equivalent prior specification for informative priors under a sequence of conditional and corresponding joint models. This additional investigation allows the imputer to perform imputations under FCS even if they only collect the prior joint information for the incomplete dataset.

For the initial step to evaluate convergence properties of FCS with informative priors, it is sensible to focus on the Bayesian normal linear models and the typical informative prior: normal inverse-gamma prior. This paper will briefly overview joint modeling, fully conditional specification, compatibility, and non-informative margins. Then, We derive a theoretical result and perform a simulation study to evaluate the non-informative margins condition. We also consider the prior for the target joint density of a sequence of normal linear models with normal inverse-gamma priors. Finally, Some remarks are concluded.

## 5.2 Background

### 5.2.1 Joint modeling

Let $Y^{obs}$ and $Y^{mis}$ denote the observed and missing data in the dataset $Y$. Joint modeling involves specifying a parametric joint model $p(Y^{obs}, Y^{mis}|\theta)$ for the complete data and an appropriate prior distribution $p(\theta)$ for the parameter $\theta$. Incomplete cases are partitioned into groups according to various missing patterns and then imputed with different submodels. Under the assumption of ignorability, the imputation model for each group is the corresponding conditional distribution derived from the assumed joint model

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|Y^{obs}, \theta)p(\theta|Y^{obs})d\theta.$$

Since the joint modeling algorithm converges to the specified multivariate distribution, once the joint imputation model is correctly specified, results will be valid and theoretical properties are satisfactory.

### 5.2.2 Fully conditional specification

Fully conditional specification attempts to define the joint distribution $p(Y^{obs}, Y^{mis}|\theta)$ by positing a univariate imputation model for each partially observed variable. The imputation model is typically a generalized linear model selected based on the nature of the missing variable (e.g. continuous, semi-continuous, categorical and count). Starting from some simple imputation methods, such as mean imputation or a random draw from the sampled values, FCS algorithms iteratively repeat imputations over all missing variables. Precisely, the $t$th iteration for the incomplete variable $Y_j^{mis}$

consists of the following draws:

$$
\begin{aligned}
\theta_j^t &\sim f(\theta_j) f(Y_j^{obs}|Y_{-j}^{t-1}, \theta_j) \\
Y_j^{mis(t)} &\sim f(Y_j^{mis}|Y_j^{obs,}Y_{-j}^{t}, \theta_j^t),
\end{aligned}
$$

where $f(\theta_j)$ is generally specified with a noninformative prior. After a sufficient number of iterations, typically ranging from 5 to 10 iterations [van Buuren, 2018, Oberman et al., 2020], the stationary distribution is achieved. The final iteration generates a single imputed dataset and the multiple imputations are created by applying FCS in parallel $m$ times with different seeds. If the underlying joint distribution defined by separate conditional models exists, the algorithm is equivalent to a Gibbs sampler.

The attractive feature of fully conditional specification is the flexibility of model specification, which allows models to preserve features in the data, such as skip patterns, incorporating constraints and logical, and consistent bounds [Van Buuren, 2007]. Such restrictions would be difficult to formulate when applying joint modeling. One could conveniently construct a sequence of conditional models and avoid the specification of a parametric multivariate distribution, which may not be appropriate for the data in practice.

### 5.2.3 Compatibility

The definition of compatibility is given by Liu et al. [2014]: let $Y = (Y_1, Y_2, \ldots, Y_p)$ be a vector of random variables and $Y_{-j} = (Y_1, Y_2, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$. A set of conditional models $\{f_j(Y_j|Y_{-j}, \theta_j) : \theta_j \in \Theta_j, j = 1, 2, \ldots, p\}$ is said to be compatible if there exists a joint model $\{f(Y|\theta) : \theta \in \Theta\}$ and a collection of surjective maps $\{t_j : \Theta \to \Theta_j\}$ such that for each $j$, $\theta_j \in \Theta_j$ and $\theta \in t_j^{-1}(\theta_j) = \{\theta : t_j(\theta) = \theta_j\}$. In that case

$$
f_j(Y_j|Y_{-j}, \theta_j) = f(Y_j|Y_{-j}, \theta).
$$

Otherwise, $\{f_j, j = 1, 2, \ldots, p\}$ is said to be incompatible. A simple example of compatible models is a set of normal linear models for a vector of continous data:

$$Y_j = N((\mathbf{1}, Y_{-j})\beta_j, \sigma_j^2),$$

where $\beta_j$ is the vector of coefficients and $\mathbf{1}$ is a vector of ones. In such a case, the joint model of $(Y_1, Y_2, \ldots, Y_p)$ would be a multivariate normal distribution and the map $t_j$ is derived by conditional multivariate normal formula. On the other hand, the classical example for an incompatible model would be the linear model with squared terms [Liu et al., 2014, Bartlett et al., 2015].

Incompatibility is a theoretical weakness of fully conditional specification since, in some cases, it is unclear whether the algorithm indeed converges to the desired multivariate distribution [Arnold and Press, 1989, Arnold et al., 2004, Heckerman et al., 2000, Van Buuren et al., 2006]. Consideration of compatibility is significant when the multivariate density is of scientific interest. Both Hughes et al. [2014] and Liu et al. [2014] stated the necessity of model compatibility for the algorithm to converge to a joint distribution. Several papers introduced some cases in which FCS models are compatible with joint distributions [van Buuren, 2018, Raghunathan et al., 2001]. Van Buuren et al. [2006] also performed some simulation studies of fully conditional specification with strongly incompatible models and concluded the effects of incompatibility are negligible. However, further work is necessary to investigate the adverse effects of incompatibility in more general scenarios.

### 5.2.4 Non-informative margins

Hughes et al. [2014] showed that the non-informative margins condition is sufficient for fully conditional specification to converge to a multivariate distribution. Suppose $\pi(\theta_j)$ is the prior distribution of the conditional model $p(Y_j|Y_{-j}, \theta_j)$ and $\pi(\theta_{-j})$ is the prior distribution of the marginal model $p(Y_{-j}|\theta_{-j})$, then the non-informative margins condition is satisfied if the joint prior could be factorized into independent priors $\pi(\theta_j, \theta_{-j}) =$

$\pi(\theta_j)\pi(\theta_{-j})$. It is worthwhile to note that the non-informative margin condition does not hold if $p(Y_j|Y_{-j},\theta_j)$ and $p(Y_{-j}|\theta_{-j})$ have the same parameter space. When the non-informative margins condition is violated, an order effect appears. In such a case, the inference of parameters would have systematic differences depending on the sequence of the variables in FCS algorithm. Simulations performed by Hughes et al. [2014] demonstrated that such an order effect is subtle. However, more research is needed to verify such claims, and it is necessary to be aware of the existence of the order effect.

## 5.3 Theoretical results

This section proves the convergence of fully conditional specification under the normal linear model with normal inverse-gamma priors to a joint distribution. Since the compatibility of the normal linear model is well understood, we will check the satisfaction of the non-informative margins condition.

Starting with the problem of Bayesian inference for $\theta = (\mu, \Sigma)$ under a multivariate normal model, let us apply the following prior distribution. Suppose that, given $\Sigma$, the prior distribution of $\mu$ is assumed to be the conditionally multivariate normal,

$$\mu|\Sigma \sim N(\mu_0, \tau^{-1}\Sigma), \tag{5.1}$$

where the hyperparameters $\mu_0 \in \mathcal{R}^p$ and $\tau > 0$ are fixed and known and where $p$ denotes the number of variables. Moreover, suppose that the prior distribution of $\Sigma$ is an inverse-Wishart,

$$\Sigma \sim W^{-1}(m, \Lambda) \tag{5.2}$$

for fixed hyperparameters $m \geq p$ and $\Lambda$. The prior density for $\theta$ can then be written as

$$\begin{aligned} \pi(\theta) \propto \ & |\Sigma|^{-(\frac{m+p+2}{2})} \ \exp\{-\tfrac{1}{2}tr(\Lambda^{-1}\Sigma^{-1})\} \\ & \times \ \exp\{-\tfrac{\tau}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\} \end{aligned} \tag{5.3}$$

For each variable $Y_j$, we partition the mean vector $\mu$ as $(\mu_j, \mu_{-j})^T$ and the covariance matrix $\Sigma$ as

$$\begin{pmatrix} \omega_j & \xi_j^T \\ \xi_j & \Sigma_{-j} \end{pmatrix},$$

such that $Y_j \sim \mathcal{N}(\mu_j, \omega_j)$ and $Y_{-j} \sim \mathcal{N}(\mu_{-j}, \Sigma_{-j})$. Similarly, we partition the scale parameter $\mu_0$ as $(\mu_{0j}, \mu_{0-j})^T$ and $\Lambda$ as

$$\begin{pmatrix} \Lambda_j & \psi_j^T \\ \psi_j & \Lambda_{-j} \end{pmatrix}.$$

The conditional model of $Y_j$ given $Y_{-j}$ is the normal linear regression $Y_j = \alpha_j + \beta_j^T Y_{-j} + \sigma_j$, where $\beta_j^T = \xi_j^T \Sigma_{-j}^{-1}$, $\alpha_j = \mu_j - \xi_j^T \Sigma_{-j}^{-1} \mu_{-j}$ and $\sigma_j = \omega_j - \xi_j^T \Sigma_{-j}^{-1} \xi_j$. The corresponding vectors of parameters $\theta_j$ and $\theta_{-j}$ would be

$$\begin{aligned} \theta_j &= (\alpha_j, \beta_j, \sigma_j) \\ \theta_{-j} &= (\mu_{-j}, \Sigma_{-j}). \end{aligned} \tag{5.4}$$

By applying the partition function illustrated by Eaton [2007, p. 165] and by block diagonalization of a partitioned matrix, the joint prior for $\theta_j$ and $\theta_{-j}$ can be derived from $\pi(\theta)$ as :

$$\begin{aligned} &\pi(\theta_j, \theta_{-j}) = p(\sigma_j) p(\beta_j | \sigma_j) p(\Sigma_{-j}) \\ &\times \exp\,\{-\tfrac{\tau}{2}(\alpha_j + \beta_j \mu_{0-j} \ - \mu_{0j})^T (\sigma_j)^{-1} (\alpha_j + \beta_j \mu_{0-j} \ - \mu_{0j})\} \\ &\times \exp\{-\tfrac{\tau}{2}(\mu_{-j} - \mu_{0-j})^T \Sigma_{-j}^{-1} (\mu_{-j} - \mu_{0-j})\} \times |\Sigma_{-j}| \\ &= \pi(\theta_j) \pi(\theta_{-j}), \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} &\pi(\theta_j) = p(\sigma_j) p(\beta_j | \sigma_j) \\ &\times \exp\,\{-\frac{\tau}{2}(\alpha_j + \beta_j \mu_{0-j} \ - \mu_{0j})^T (\sigma_j)^{-1} (\alpha_j + \beta_j \mu_{0-j} \ - \mu_{0j})\}, \end{aligned} \tag{5.6}$$

$$\pi(\theta_{-j}) = p(\Sigma_{-j}) \times \exp\{-\frac{\tau}{2}(\mu_{-j} - \mu_{0-j})^T \Sigma_{-j}^{-1} (\mu_{-j} - \mu_{0-j})\} \times |\Sigma_{-j}| \tag{5.7}$$

and

$p(\sigma_j) \sim W^{-1}(m, \lambda_j)$, $p(\beta_j|\sigma_j) \sim \mathcal{N}(\psi_j^T \Lambda_{-j}^{-1}, \lambda_j \Lambda_{-j}^{-1})$, $p(\Sigma_{-j}) \sim W^{-1}(m-1, \Lambda_{-j})$, $\lambda_j = \Lambda_j - \psi_j^T \Lambda_{-j}^{-1} \psi_j$ (Eaton, 2007, Section 8.2). Since the joint prior distribution factorizes into independent priors, the "non-informative" margins condition is satisfied. Based on equations (6) and (7), we could derive the prior for the conditional linear model from the prior for the multivariate distribution:

$$\begin{aligned} &p(\sigma_j) \sim W^{-1}(m, \lambda_j) \\ &p(\beta_j|\sigma_j) \sim \mathcal{N}(\psi_j^T \Lambda_{-j}, \lambda_j \Lambda_{-j}) \\ &p(\alpha_j|\sigma_j) \sim \mathcal{N}(\mu_{0j} - \psi_j^T \Lambda_{-j} \mu_{02}, \tau^{-1}\sigma_j - (\mu_{-0j})^2 \lambda_j \Lambda_{-j}^{-1}) \end{aligned} \tag{5.8}$$

Since the conditional $\beta_j|\sigma_j$ follows a normal distribution, the marginal distribution $\beta_j$ would be a student's t-distribution $\beta_j \sim t(\psi_j^T \Lambda_{-j}^{-1}$, $m\Lambda_{-j}^{-1}\lambda_j^{-1}, 2m - p + 1)$. When the sample size increases, $\beta_j$ tends to the normal distribution $N(\psi_j^T \Lambda_{-j}^{-1}, \frac{\lambda_j \Lambda_{-j}}{m-1})$. Similarly, the marginal distribution $\alpha_j$ would be $t(\mu_{0j} - \psi_j^T \Lambda_{-j} \mu_{02}, m(\tau^{-1} - (\mu_{0-j})^2 \Lambda_{-j}^{-1}) \Lambda_j^{-1}, 2m - p + 1)$. When the sample size increases, $\alpha_j$ tends to the normal distribution $N(\mu_{0j} - \psi_j^T \Lambda_{-j} \mu_{02}$, $\frac{1}{(\tau^{-1} - (\mu_{-0j})^2 \Lambda_{-j}^{-1})(m-1)} \Lambda_j)$. Usually, when the sample size is over 30, the difference between student's t-distribution and the corresponding normally distributed approximation is negligble. With the prior transformation formula, one could apply Bayesian imputation under the normal linear model with normal inverse-gamma priors. This holds for both the prior information about the distribution of the data (e.g. location and scale of variables) and the scientific model (e.g. regression coefficients).

## 5.4 Simulation

We perform a simulation study to demonstrate the validity and the convergence of fully conditional specification when the conditional models are simple linear regressions with an inverse gamma prior for the error term and a multivariate normal prior for regression weights. In addition, we look for the disappearance of order effects, which is evident in the convergence of fully conditional specification to a multivariate distribution.

We repeat the simulation 500 times and generate a dataset with 200 cases for every simulation according to the following multivariate distribution :

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 1 \\ 4 \\ 9 \end{pmatrix}, \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 9 \end{pmatrix}\right]$$

Fifty percent missingness is induced on either variable $x$, $y$ or $z$. The proportion of the three missing patterns is equal. When evaluating whether it is appropriate to specify a normal inverse gamma prior, we consider both missing completely at random (MCAR) mechanisms and right-tailed missing at random (MARr) mechanisms where higher values have a larger probability to be unobserved. When investigating the existence of order effects, we only conduct the simulation under MCAR missingness mechanism to ensure that the missingness does not attribute to any order effects. We specify a weak informative prior for two reasons. First, with a weak informative prior, the frequentist inference is still plausible by applying Rubin's rules [Rubin, 1987, p. 76]. Second, Goodrich et al. [2019] suggested that compared with flat non-informative priors, weak informative priors places warranted weight to extreme parameter values. In such a case, The prior under the joint model is specified as: $\mu_0 = (0, 0, 0)^T$, $\tau = 1$, $m = 3$ and

$$\Lambda = \begin{pmatrix} 60 & 0 & 0 \\ 0 & 60 & 0 \\ 0 & 0 & 60 \end{pmatrix}$$

and the corresponding prior for separated linear regression model would be the same, with $\pi(\sigma) \sim W^{-1}(3, 60)$ and

$$(\alpha, \beta)^T \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 60 & 0 & 0 \\ 0 & 3600 & 0 \\ 0 & 0 & 3600 \end{pmatrix}\right].$$

### 5.4.1 Scalar inference for the mean of variable Y

The aim is to assess whether Bayesian imputation under a normal linear model with normal inverse gamma priors would yield unbiased estimates and exact coverage of the nominal 95% confidence intervals. Table 5.1 shows that with weak informative prior, fully conditional specification also provides valid imputations. The estimates are unbiased, and the coverage of the nominal 95% confidence intervals is correct under both MCAR and MARr. Without the validity of a normal inverse gamma prior specification, further investigations into the convergence would be redundant.

| | Bias | Cov | Ciw |
|---|---|---|---|
| MCAR | 0 | 0.95 | 0.74 |
| MARr | -0.01 | 0.97 | 0.73 |

Table 5.1: Bias of the estimates ($E(Y)$) and coverage of nominal 95% confidence intervals under MCAR and MARr

### 5.4.2 order effect evaluation

The visit sequence laid upon the simulation is $z$, $x$ and $y$. To identify the presence of any systematic order effect, we estimate the regression coefficient directly after updating variable $z$ and after updating variable $x$. Specifically, the $i$th iteration of fully conditional specification would be augmented as:

1. impute $z$ given $x^{i-1}$ and $y^{i-1}$.

2. build the linear regression $y = \alpha + \beta_1 x + \beta_2 z + \epsilon$ and collect the coefficient $\beta_1$, donoted as $\hat{\beta_1}^z$.

3. impute $x$ given $z^i$ and $y^{i-1}$.

4. build the linear regression $y = \alpha + \beta_1 x + \beta_2 z + \epsilon$ and collect the coefficient $\beta_1$, donoted as $\hat{\beta_1}^x$.

5. impute $y$ given $z^i$ and $x^i$.

After a burn-in period with 10 iterations, the fully conditional specification algorithm was performed with an additional 1000 iterations, in which differences between the estimates $\hat{\beta_1}^z - \hat{\beta_1}^x$ are recorded. The estimates from the first 10 iterations are omitted since the FCS algorithms commonly reach convergence around 5 to 10 iterations. Estimates from the additional 1000 iterations would be partitioned into subsequences with equal size, which are used for variance calculation. We calculate the nominal 95% confidence interval of the difference. The standard error of the difference is estimated with batch-means methods [Albert, 2009, p. 124]. The mean of $\hat{\beta_1}^z - \hat{\beta_1}^x$ is set to zero. Since only three 95% confidence intervals derived from 500 repetitions do not cross the zero, there is no indication of any order effects. We also monitor the posterior distribution of the coefficient under both joint modeling and fully conditional specification. Figure 5.1 shows a quantile-quantile plot demonstration the closeness of the posterior distribution for $\beta_1$ derived from both joint modeling and fully conditional specification. Since the posterior distributions for $\beta_1$ under joint modeling and FCS are very similar, any differences may be considered negligible in practice.

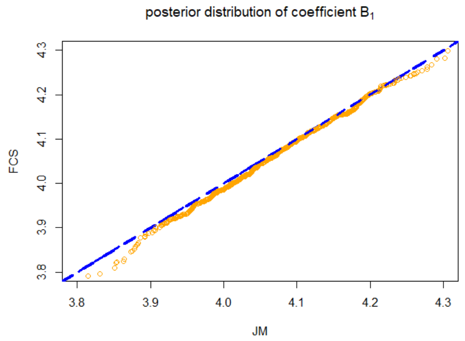


Figure 5.1: qqplot demonstrating the closeness of the posterior distribution of JM and FCS for $\beta_1$

All these results confirm that under the normal inverse gamma prior, Bayesian imputation under normal linear model converges to the corresponding multivariate normal distribution.

## 5.5 Conclusion

Based on the theory of the non-informative margins condition proposed by Hughes et al. [2014], we prove the convergence of fully conditional specification under the normal linear model with normal-inverse-gamma prior distributions. Since it has been shown that a sequence of normal linear models is compatible with a multivariate normal density, we only focus on the non-informative margins condition for the prior. The transformation of the prior between a normal inverse gamma for fully conditional specification and a normal inverse Wishart for joint modeling is useful. With transformation, one could apply fully conditional specification and only collect prior information relative to the distribution of variables rather than scientific models.

Fully conditional specification is an appealing imputation method because it allows one to specify a sequence of flexible and simple conditional models and bypass the difficulty of multivariate modeling in practice. The default prior for normal linear regression is Jeffreys prior, which satisfies the non-informative margin condition. However, it is worth developing other types of priors for fully conditional specification such that one could select the prior, which suits the description of prior knowledge best. Many researchers have discussed the convergence condition of FCS. However, there is no conclusion for the family of posterior distributions that satisfies the condition of convergence. In such a case, when including new kinds of priors in fully conditional specification algorithms, it is necessary to investigate the convergence of the algorithm with new posterior distributions. Specifically, one should study the non-informative margin conditions for new priors. Compatibility should also be considered if the imputation model is novel. Our work takes steps in this direction.

Although a series of investigations have shown that the adverse effects of violating compatibility and the non-informative margin conditions may be small, all of these investigations rely on pre-defined simulation settings. More research is needed to verify conditions under which the fully conditional specification algorithm converges to a multivariate distribution and cases in which the violation of compatibility and non-informative margin has negligible adverse impacts on the result.

There are several directions for future research. From one direction, it is possible to develop a prior setting to eliminate order effects of the fully conditional specification algorithm under the general location model since the compatibility and non-informative margins conditions are satisfied under the saturated multinomial distribution. Moreover, various types of priors of the generalized linear model for the fully conditional specification and corresponding joint modeling rationales could be developed. Another open problem is the convergence condition and properties of block imputation, which partitions missing variables into several blocks and iteratively imputes blocks [van Buuren, 2018, Section 4.7.2]. Block imputation is a more flexible and user-friendly method. However, its properties have yet to be studied. Finally, it is necessary to investigate the implementation of prior specifications in software.

# Part III

# Diagnostics for imputation models

# Chapter 6

# Graphical and numerical diagnostic tools to assess multiple imputation models by posterior predictive checking

## Abstract

We propose a method to diagnose imputation models based on posterior predictive checking. To assess the congeniality of imputation models, we compare the observed data with their replicates generated under corresponding posterior predictive distributions. The idea is that if the imputation model is congenial with the substantive model, the observed data is expected to locate in the center of corresponding predictive posterior distributions. We investigate the proposed diagnostic method for parametric and non-parametric imputation approaches, continuous and discrete incomplete variables, univariate and multivariate missingness patterns. The results show the validity of the proposed diagnostic method.

## 6.1 Introduction

Multiple imputation (MI) is a popular approach for the analysis of incomplete datasets. It involves generating several plausible imputed datasets and aggregating different results into a single inference. Missing cells are filled with synthetic data drawn from corresponding posterior predictive distributions. This procedure is repeated multiple times, resulting in several imputed datasets. The scientific interests are then estimated for each imputed dataset by complete-data analyses. Finally, different estimates are pooled into one inference using Rubin's rule, which accounts for within and across imputation uncertainty [Rubin, 1987].

A crucial part of the multiple imputation process is selecting sensible models to generate plausible values for the missing data. The validity of post-imputation analyses relies on the congeniality of the imputation model and the substantive model of interest [Meng, 1994]. However, the model selection is not a trivial model selection process in practice since there can be a wide array of candidate models to check. Researchers should consider which variables, interaction terms, and nonlinear terms are included based on the scientific interest and the data features.

Despite the popularity of multiple imputation, there are only a few imputation model diagnostic methodologies. One standard diagnostic method is to compare distributions of the observed data and the imputed data [van Buuren, 2018, Abayomi et al., 2008]. Plausible imputation models would generate imputed values that have a similar distribution to the observed data. Although missing at random (MAR) mechanisms would also induce the discrepancies between the observed and imputed data, any dramatic departures that the observed data features cannot explain are evidence of potential imputation model misspecification. Reliable interpretation of the observed and imputed data's discrepancies could be derived from external knowledge about the incomplete variables and the missingness mechanisms [Abayomi et al., 2008].

The idea to evaluate the validity of scientific models on multiply imputed data is not new. Bondarenko and Raghunathan [2016] proposed an advanced diagnostic method to compare the distributions of the observed and imputed data conditional on the probability of missingness. Gelman et al. [1998] applied cross-validation to check the fit of a hierarchical Bayesian model in the study of 51 public opinion polls preceding the 1988 U.S. Presidential election. Gelman et al. [2005] also proposed to apply graphical posterior predictive inference on the test statistics for model checking with missing and latent data. If regression-based imputation approaches are involved, the conventional regression diagnostics, such as plots of residuals and outliers, are helpful [White et al., 2011]. A comprehensive overview of model diagnostic in multiple imputation is available in Nguyen et al. [2017].

Posterior predictive checking (PPC) has been proposed as an alternative method for the imputation model diagnostic [Gelman et al., 2005, He and Zaslavsky, 2012, Nguyen et al., 2015]. PPC is a Bayesian model checking approach that compares the replicated data drawn from the corresponding posterior predictive distribution to the observed data. If the model lacks fit, there could be a discrepancy between the replicated and observed data.

He and Zaslavsky [2012] and Nguyen et al. [2015] applied posterior predictive checking to assess the inadequacies of the joint imputation model with one or more test quantities relative to the scientific interest. To evaluate the 'usability' of imputation model with respect to the test statistics, analysts compares the estimates for the complete data and their replicates, $Q(Y_{obs}, Y_{mis} - Q(Y_{com}^{rep}))$. Comparisons of the complete data and its replicates ensure the calculation of test quantities with general missingness patterns. However, it also results in sensitivity to the amount of missingness.

This manuscript proposes and evaluates the implementation of posterior predictive checking for imputation techniques that rely on the observed data. The general idea is that if the imputation model is congenial with the substantive model, the observed data is expected to locate in the center of corresponding predictive posterior distributions. We compare the completed data and their posterior replicates simulated under the model and evaluate the location of the observed data. This distinguishes our approach from the posterior predictive checking of imputation models by applying target analyses. We demonstrate:

1. that PPC can be generalized to variable-by-variable imputation techniques;
2. that PPC can be used to identify the imputation model that conforms most to the true data generating model;
3. that PPC can be used as a model evaluation technique to identify the better substantive analysis model;
4. how to perform PPC with `MICE` in `R` on a real-life data set [van Buuren and Groothuis-Oudshoorn, 2011];
5. that this PPC approach is not sensitive to the amount of missing data.

The remainder of this manuscript is organized as follows. In section 6.2, we review the posterior predictive checking of the imputation model by applying target analysis proposed by He and Zaslavsky [2012]. In section 6.3, we provide an overview of the `MICE` package and the underlying imputation algorithm: fully conditional specification (FCS). We also further point out the necessity of extending the posterior predictive checking of the imputation model so that the diagnostics would apply to the `MICE` algorithm. In section 6.4, we evaluate the performance of the proposed diagnostic approach with simulation studies. In section 6.5, we show the results of simulation studies. In section 6.6, we apply the proposed diagnostic approach to the body mass index (BMI) data in Dutch. In section 6.7, we conclude with a discussion of our findings.

## 6.2 Posterior predictive checking (PPC)

### 6.2.1 Posterior predictive checking

Without incomplete variables, PPC compares the observed data $y$ with the replicated data $y^{rep}$ which are simulated from the posterior predictive distribution, with parameter $\theta$:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta \tag{6.1}$$

To detect the discrepancy between the model and the data, we define test quantities which reflect the scientific interest and estimate them for both observed data and replicated data. Misfit of the model with respect to the data could be summarized by the posterior predictive p-value, which is the probability that the replicated data are more extreme than the observed data, with respect to the selected test quantities [Gelman et al., 2013]:

$$\begin{aligned} p_B &= Pr(T(y^{rep}, \theta) \geq T(y, \theta) | y) \\ &= \int \int I_{T(y^{rep}, \theta) \geq T(y, \theta)} p(y^{rep} | \theta) p(\theta | y) dy^{rep} d\theta, \end{aligned} \tag{6.2}$$

where $I$ is the indicator function. An extreme p-value (close to 0 or 1) implies the suspicion on the fit of the model since a consistent discrepancy between test quantities $T(y^{rep}, \theta)$ and $T(y, \theta)$ cannot explained by simulation variance.

Posterior predictive checking has been widely used for model diagnostic in applied Bayesian analysis [Gelman et al., 2013, chapter 6], and the posterior predictive distribution is usually calculated by simulation. Suppose we have $N$ draws of model parameters from its posterior distribution $\theta_j, j = 1, \ldots, N$, we then generate a replicated data for every theta $\theta_j$. The PPC compares test quantities based on observed data with the empirical predictive distribution of test quantities. The estimated posterior predictive p-value is the proportion of these N simulations for which $T_j(y^{rep}, \theta) > T_j(y, \theta)$. It is noticeable that PPC's application for the imputation model diagnostic is not based on the hypothesis test perspective. Hence there is no underlying assumed distribution for the posterior predictive p-value in this case. The posterior predictive p-value indicates whether the model would provide plausible inference based on the data with respect to the selected test quantities.

To perform multiple imputation model checking with PPC, we compare the completed data, the combination of the observed and imputed data, with its replications. Gelman et al. [2005] applied graphical PPC to visualize test quantities comparisons based on completed and replicated data. He and Zaslavsky [2012] and Nguyen et al. [2015] developed numerical posterior predictive checks as target analyses to the joint imputation model. He and Zaslavsky [2012] proposed two kinds of discrepancies, completed data discrepancy and expected completed-data discrepancy, and the approaches to calculate corresponding posterior predictive p-values. We briefly introduced these discrepancies and p-values for the completeness of PPC for MI models.

### 6.2.2 Complete data discrepancy

To assess the complete data discrepancy $T(y_{com}^{rep}, \theta) - T(y_{com}, \theta)$, we draw imputed values for incomplete variables $y_{mis}$ and the replication of the complete data $y_{com}^{rep}$ from their posterior predictive distribution:

$$p(y_{com}^{rep}, y_{mis}|y_{obs}) = \int p(y_{com}^{rep}|\theta)p(y_{mis}|y_{obs}, \theta)p(\theta|y_{obs})d\theta, \tag{6.3}$$

where $y_{obs}$ is the observed data. To assess the model fit, we calculate the posterior predictive p value as :

$$\begin{aligned} p_{B,com} &= Pr(T(y_{com}^{rep}) \geq T(y_{com})|y_{obs}) \\ &= \int\int I_{T(y_{com}^{rep}) \geq T(y_{com})} p(y_{com}^{rep}, y_{mis}|y_{obs}) dy_{com}^{rep} dy_{mis} \end{aligned} \tag{6.4}$$

The simulation process to estimate p-value proposed by He and Zaslavsky [2012] is:

1. Simulate $N$ draws of $\theta$ from the corresponding posterior distribution $p(\theta|y_{obs})$
2. For each $\theta_j, j = 1, \ldots, N$, impute $y_{mis}^j$ from $p(y_{mis}|y_{obs}, \theta_j)$ and simulate the replicated data $y_{com}^{rep,j}$ from $p(y_{com}^{rep}|\theta_j)$

A $p_{B,com}$, which is close to 0 or 1, implies the discrepancy between the model and the data with respect to the selected test quantities.

### 6.2.3 Expected complete data discrepancy

He and Zaslavsky [2012] noticed that the power of complete data discrepancy is weakened because the variance of imputed data across complete data $y_{imp}$ and replicated data $y_{imp}^{rep}$ increase the variance of the test quantities. He and Zaslavsky [2012] reduced the variance of complete data discrepancy by calculating the expectation value of missing data for each model parameter draw. The modification of p-value $p_{B,ecom}$ would be:

$$\begin{aligned} p_{B,ecom} &= Pr(E[T(y_{com}^{rep})|y_{obs}^{rep}, y_{obs}] \geq E[T(y_{com})|y_{obs}^{rep}, y_{obs}]|y_{obs}) \\ &= \int\int I_{E[T(y_{com}^{rep})|y_{obs}^{rep}, y_{obs}] \geq E[T(y_{com})|y_{obs}^{rep}, y_{obs}]} p(y_{obs}^{rep}, y_{obs}) dy_{obs}^{rep} \end{aligned} \tag{6.5}$$

Again, the nested simulation process to calculate the p-value $p_{B,ecom}$ is:

1. Simulate $N_1$ draws of $\theta$ from the corresponding posterior distribution $p(\theta|y_{obs})$

2. For each $\theta_j, j = 1, \ldots, N_1$, impute $y_{mis}^j$ from $p(y_{mis}|y_{obs}, \theta_j)$ and simulate the replicated data $y_{com}^{rep,j}$ from $p(y_{com}^{rep}|\theta_j)$

3. For each j-th replicate, calculate the mean discrepancy by setting $y_{mis}^j$ and $y_{com}^{rep,j}$ to missing and overimputing them with the same paramters $\theta_j$ over $N_2$ draws $y_{mis}^{j,k}$ and $y_{com}^{rep,j,k}, k = 1, \ldots, N_2$. Calculate the difference : $D_{j,k} = T(y_{obs}^{rep,j}, y_{mis}^{rep,j,k}) - T(y_{obs}, y_{mis}^{rep,j,k})$ over $N_2$ draws and then average the difference for the j-th replicate : $\bar{D}_{j.} = \sum_1^k D_{j,k}/k$

4. Calculate $p_{B,ecom}$ as the proportion of these $N_1$ estimates that are positive, $\bar{D}_{j.} \geq 0$

He and Zaslavsky [2012] evaluated whether the PPC could detect the uncongeniality of the imputation model. Nguyen et al. [2015] investigated the performance of PPC in other imputation model misspecification scenarios, such as ignoring the response variable and auxiliary variables or failing to transform skewed variables. The PPC approach proposed by He and Zaslavsky [2012] is based on the joint imputation model. The imputation model for diagnostic is a joint distribution for the observed data, and the test quantities depend on multiple variables and parameters.

## 6.3 `MICE` package

There are two popular approaches for multiple imputation: joint modeling (JM) and fully conditional specification (FCS). JM involves specifying a multivariate distribution for the missing data and generating imputations from the conditional distribution based on different missing patterns. For example, suppose that for an incomplete dataset $Y = (Y_1, Y_2, Y_3, Y_4)$ the missing pattern for case $i$ is $r_{[i]} = (0, 0, 1, 1)$, where 0 denotes missing and 1 denotes being observed, JM specifies a joint modeling for these four variables and impute the case $i$ with the bivariate conditional distribution $P_i(Y_1^{mis}, Y_2^{mis}|Y_3^{obs}, Y_4^{obs}, \theta_{1,2})$.

FCS attempts to specify an imputation model for each missing variable $Y_j$ conditional on all the other variables $P(Y_j|Y_{-j}, \theta_j)$, with patameter $\theta_j$. It generates imputations iteratively over all missing variables after an initial imputation, such as mean imputation or random draw from observed values. Given the most recent imputations of the other missing variables $Y_j^t$ at iteration $t$, the algorithm of generating imputations for the missing variable $Y_j$ consists of the following draws:

$$\theta_j^t \sim f(\theta_j)f(Y_j^{obs}|Y_{-j}^{t-1}, \theta_j)$$
$$Y_j^{mis(t)} \sim f(Y_j^{mis}|Y_{-j}^t, \theta_j^t),$$

where $f(\theta_j)$ is the prior distribution for the parameter of the imputation model $\theta_j$. The FCS is an attractive imputation approach because of its flexibility in imputation model specification. It is known under different names: chained equations stochastic relaxation, variable-by-variable imputation, switching regression, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC, and iterated univariate imputation [Van Buuren, 2007].

Multivariate Imputation by Chained Equations (`MICE`) is the name of software for imputing incomplete multivariate data by Fully Conditional Specification (FCS). It has developed into the de facto standard for imputation in `R` and is increasingly being adopted in Python (e.g., statsmodels (imputer function) & miceforest). The `MICE` package creates functions for three components of FCS: imputation, analysis, and pooling. Figure 6.1 illustrates how MICE solves a missing data problem by generating 3 imputed datasets. Three imputed datasets are generated with function **mice()**. Analysis are performed on every imputed dataset by **with()** function and combined into a single inference with function **pool()**. The software stores the output of each step in a particular class: **mids**,

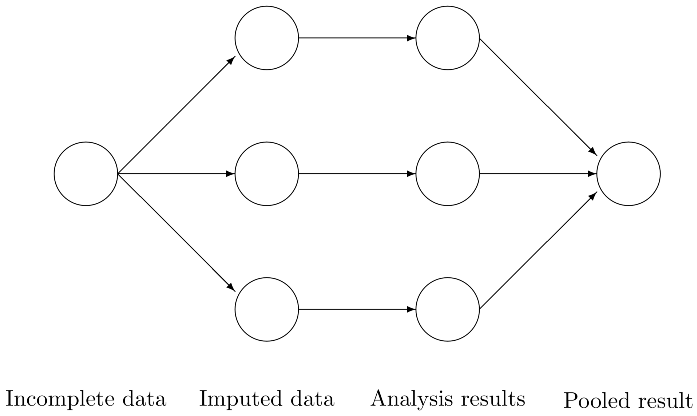


Figure 6.1: Main steps used in `MICE` [van Buuren and Groothuis-Oudshoorn, 2011]

**mira** and **mipo**. More details aboout `MICE` package can be found in van Buuren and Groothuis-Oudshoorn [2011].

Two features of the software motivate our research. First, the default imputation method for numerical missing data is predictive mean matching (PMM) [Little, 1988].

```r
library(mice, warn.conflicts = FALSE)
imp <- mice(nhanes, print = FALSE)
imp$method
```

```
##    age   bmi   hyp   chl
##     ""  "pmm" "pmm" "pmm"
```

It generates imputations for a missing cell from its p nearest points. The distance function applied to selection nearest points in MICE is:

$$d_{mice}(x_i^{obs}, x_j^{mis}) = |x_i^{obs}\beta^* - x_j^{mis}\hat{\beta}|, \tag{6.6}$$

where $\hat{\beta}$ is the mean of posterior distribution of models' parameters and $\beta^*$ is a random draw from the corresponding posterior distribution. Predictive mean matching is a nonparametric imputation approach which is proven to perform well in a wide range of scenarios [De Waal et al., 2011, Siddique and Belin, 2008, Su et al., 2011, van Buuren, 2018, van Buuren and Groothuis-Oudshoorn, 2011, Vink et al., 2014, White et al., 2011, Vink et al., 2015, Yu et al., 2007]. The attractive advantage of PMM is that the imputed data is consistently within the range of the sample space [Heeringa, 2001, van Buuren, 2018, Vink et al., 2014,0, White et al., 2011, Yu et al., 2007]. For instance, PMM prevents imputing negative values for data that are strictly non-negative. Second, **mids** only stores imputed datasets not the estimated parameters of the imputation models [Hoogland et al., 2020].

Based on the features of `MICE` package discussed above, it is necessary to investigate whether PPC could check the donor selection procedure of PMM and perform PPC based on the observed data itself instead of the target statistics. He and Zaslavsky [2012] briefly discussed the approach to checking imputation models for subsets of missing variables. However, they assumed that the imputations of the remaining variables (excluding the incomplete variables of interest in an assessment) are adequate. We also evaluate the performance of PPC when relaxing this assumption in the application section 6.6.

The implement of PPC for multiple imputation models in `MICE` package is straightforward. Suppose we would like to diagnose the congeniality of the imputation model for a block of missing variables $Y_{interest}$, we duplicate the complete cases of $Y_{interest,obs}$ and remove the values of $Y_{interest}$ in the duplication part $Y^{dup,mis}_{interest,obs}$. It is noticeable that we discard the cases with missing cells in $Y_{interest}$ in the original dataset. The derived concatenated dataset has two parts of the same cases. One half is the cases with completely observed $Y_{interest}$ in both original data and the concatenated data. The other half is the cases with completely observed $Y_{interest}$ in original data but removed the observed value deliberately in the concatenated data. The remaining variables could also be partially observed. To perform the PPC for the imputation model of $Y_{interest}$, we apply the `mice` function in the `MICE` package to impute the concatenated data. Imputation values for $Y^{dup,mis}_{interest,obs}$ in a number of iterations forms the posterior distribution for each case with completely observed $Y_{interest}$. To derive stable posterior distribution, the number of iterations *nit* should not be too small. He and Zaslavsky [2012] suggested it is sufficient to set *nit* from 20 to 50.

## 6.4 Simulation Study

We carried out a simulation study to illustrate the performance of the proposed diagnostic approach, varying several factors including missingness proportion (30%, 50%, 80%), missingness mechanisms (MCAR and MARr), nominal levels of the confidence interval (75%, 95%) and different imputation models. The simulation study consisted of diagnostics under three scenarios: 1) quadratic equation with an incomplete outcome 2) quadratic equation with missing covariates 3) generalized linear model with an incomplete binary outcome. We evaluated whether the proposed diagnostic method could identify the congeniality imputation model for continuous and discrete missing variables under the first and the third scenarios. We also investigated the performance of the proposed diagnostic method on the donor selection procedure of predictive mean matching under the second scenario. The sample size and the number of iterations were set to be 1000 and 50 separately in all simulations.

We induced missingness with the `ampute()` function in the simulation study. Generally, `ampute()` is a convenient function in `MICE` package to generate missing data for simulation purposes [Schouten et al., 2018]. We considered missing completely at random (MCAR) mechanism where the probability of missing is equal for every cell as well as right-tailed missing at random (MARr) mechanism where higher values of covariates have a higher probability of being unobserved. In the algorithm of `ampute()` function, the probability of missing is allocated with different logistic functions of weighted sum score, which is a linear combination of covariates correlated with the probability of missing:

$$wss_i = w_i x_{1i} + w_i x_{2i} + \cdots + w_i x_{mi} \tag{6.7}$$

The weight $w_i$ is pre-specified to reflect the influence of the variable $x_i$ on the probability of missing. For instance, if the formation of a weighted sum score is: $wss = x_1 + x_2$, the probability of missing is determined by both $x_1$ and $x_2$ with the equal effects. More specifically, under MARr mechanism, candidates with higher values of weighted sum score have a higher probability of being unobserved when applying the `ampute()` function to generate missing data.

### 6.4.1 Quadratic equation with an incomplete outcome

In the first simulation study, we considered a partially observed variable $Y$ and a fully observed variable $X$. The data was generated from : $X \sim unif(-3, 3)$, $Y|X \sim \mathcal{N}(X + X^2, 1)$. The scientific model was indeed a quadratic model. We considered two imputation models for the missing response $Y$: one is a linear regression of $Y$ on $X$, and the other is a quadratic regression of $Y$ on $X$.

### 6.4.2 Quadratic equation with incomplete covariates

In the second simulation study, the response variable $Y$ was generated from a normal distribution: $Y|X \sim \mathcal{N}(X + X^2, 1)$, where the covariate $X$ followed a standard normal distribution. In this simulation study, the response variable $Y$ was completely observed while the covariate $X$ and the corresponding quadratic term $X^2$ were jointly missing for a part of cases. There were no cases with missing cells on either $X$ or $X^2$. We compared two non-parametric methods, predictive mean matching (PMM) and polynomial combination (PC) with a parametric method, the substantive model compatible fully conditional specification (SMC-FCS). The PC and SMC-FCS methods are two accepted methods to impute linear regression with quadratic terms. The PC method is an extension of PMM but applies a different donors selection procedure.

### 6.4.3 Generalized linear model for discrete variables

The final simulation study considered a partially observed binary $Y$ and two complete covariates $X$ and $Z$. The model of scientific interest was : $Pr(Y = 1|X, Z) = exp(X + Z)/1 + exp(X + Z)$, where $x \sim unif(-3, 3)$ and $Z \sim \mathcal{N}(1, 1)$. Under MARr mechanism, the weights of variables $X$ and $Z$ in the determining of the probability of missing for $Y$ were set to be equal. Since the logistic regression actually models the probability of assignment, we investigated the plot of deviance and calculated the sum of squared deviance divided by the sample size. There were two candidate models: a logistic regression of $Y$ on $X$ and $Z$ and a logistic regression of $Y$ on $Z$ only.

## 6.5 Simulation results

In this section, we present the simulation results of the proposed diagnostic method under three different scenarios. For numerical assessment, We estimated the rate by which the nominal confidence interval covers the observed data (COV), the mean of the distance between the observed data and the location of the predictive posterior distribution (Distance), and the average width of the nominal confidence intervals (CIW). We also provided graphical analysis with scatterplots, density plots and distribution plots, which show the means and nominal confidence intervals for each observed value.

### 6.5.1 Quadratic equation with an incomplete outcome

Table 6.1 shows the result of the simulation study when the substantive model is a quadratic equation with an incomplete outcome. Since we only generated one incomplete dataset and repeat imputing it 50 times, all coverage rates were close to the pre-specified nominal level. However, when the imputation model was misspecified as a linear regression model, the average distance was larger than the average distance under the correct specification of the imputation model (linear regression with a quadratic term). It proved our intuitive idea that the data would be close to the center of predictive posterior distributions if the model fits. The variance of the missing variable $Y$ was set as 1, which implied that the width of 95% nominal confidence interval is approximate 3.92 (1.96 * 2) and the width of 75% nominal confidence interval is approximate 2.3 (1.15 * 2). When the imputation model was correctly specified, the estimated average width of the confidence interval was unbiased. However, the variance of $Y$ was overestimated when the imputation model was linear.

The same result could also be derived from the graphical analysis. Figure 6.2 shows means and 95% confidence intervals (distribution plot) under the scenario of 30% missing cases and MARr missing mechanism. When the imputation model was correctly specified, the points out of the 95% confidence interval randomly spread over the sample space without any patterns. However, when the imputation was incorrectly specified as the linear regression model, the points out of the confidence interval clustered in the regions with extreme values of $X$. The density plot and the scatter plot of the observed and replicated data generated with function **densityplot()** and **xyplot()** in `MICE` also show the evidence that the quadratic regression is preferable than the linear model (see figure 6.3). The scatter plot of the quadratic regression imputation model shows that replicated data overlapped the observed data. The density plot shows that the replicated data shared the same distribution as the observed data. This evidence illustrates the congeniality of the quadratic regression imputation model. However, the linear regression model performed worse than the quadratic regression model. First, the replicated data did not cover the observed data in two extreme regions in the scatter plot. Second, the empirical density of the replicated data and observed data were different.

These three plots do not merely illustrate the misspecification of the imputation model as the linear regression model. They also provide information to identify the regions of sample space in which the sub-optimal imputation model could still generate acceptable imputed values. Based on the plot of means and 95% confidence intervals for the linear regression imputation model, we could also develop the piecewise linear regression model for the observed data.

When we cannot figure out the imputation model under which the observed data fit the predictive posterior distribution perfectly, these plots based on observed data provide the evidence of rebuild a contamination imputation model, which would improve the validity of imputation values. When the missing cases are not in the regions where outliers crowd, we could even apply the incongenial imputation model. For example, in our simulation scenarios, suppose we only consider the linear regression imputation model and missing cases with near the parabolic minimum $X$ values. The imputed value will not show significant deviations from the true value. Finally, our proposed PPC approach for imputation models is robust against the different percentages of missing cases, missingness mechanisms, and the confidence interval's nominal levels. The nominal level of the confidence interval is determined by the extent to which we could undertake the outliers when the imputation model is not congenial with the data generating process. For instance, there were more outliers in the plot of means and 75% confidence intervals than the plot of mean and 95% confidence intervals. When we would like to replace the linear regression model with a piecewise linear model to improve the imputation, the selection knots based on the 75% confidence intervals are more closed to the parabolic minimum than the 95% confidence intervals (See figure 6.4).

### 6.5.2 Quadratic equation with incomplete covariates

Table 6.2 and 6.3 show the result of the simulation for the quadratic equation with incomplete covariates. Based on the numerical result, the performance of these three methods, PC, SMC-FCS, and PMM, was the same, despite the slightly reduced coverage rate of PMM. In fact, when the missingness mechanism is MCAR (to bypass the problem of the sparse observed region for PMM), the PMM would also provide the valid inference of the regression weights (see Table 6.4) [Vink and van Buuren, 2013].

However, when it comes to graphical diagnostics, the misfit of PMM appears. The distribution plot (figure 6.5 and 6.6) show that PC and SMC-FCS generated the same posterior predictive distribution of the observed data. There were more outliers with a larger value of $Y$. It is sensible since the density function of $X$ based on $Y$ is not monotone. Thus, it is unavoidable to impute the missing cell on the opposite arm of the parabolic function. Although in such a case, the imputed value was not the same as the true value, the replicated data still overlapped the observed data in the scatter plot (see Figures 6.7). The plot of mean and 95% nominal confidence interval of PMM in Figure 6.5 did not show more outliers than these of PC and SMF-FCS. However, when the nominal level was set as 75%, more outliers appeared in the sub-plot of PMM (Figure 6.6). The reason is that there are more observed data closed to the center in the plots of PC and SMC-FCS, which implies the superiority of PC and SMC-FCS. The scatter plot also shows the discrepancy between the distribution of the replicated and the observed data with respect to PMM (Figures 6.7). The result is robust against various percentages of missing cases and missing mechanisms. The proposed PPC for imputation model could check the donors selection process of hot-deck approaches. In our simulation scenarios, selecting donors for the composed variable $X + X^2$ was better than only the missing variable $X$. The SMC-FCS was treated as the baseline in our simulation since it is proven as a reliable imputation method when the substantive model is known [Bartlett et al., 2015]. The PC performs as well as the SMC-FCS which implies the donor selection process of PC reflects the data generating process in our simulation scenarios. However, when applying hot-deck approaches to implement imputation, the success in model checking is not sufficient to valid imputations. In order to determine whether we could derive plausible imputations, additional diagnoses of distributions of the complete variable for observed and missing cases is necessary. For example, in our simulation scenarios, if the variable $Y$'s range of missing cases is out of the range of observed cases, the PC may still derive implausible imputations.

### 6.5.3 Generalized linear model for discrete variables

Table 6.5 shows the average sum of squared deviance for two different logistic regression models. The value of the average sum of squared deviance was smaller when the imputation model was correctly specified with logistic regression on both $X$ and $Z$. The result is robust against the percentage of missing cases and missingness mechanisms. Figure 6.8 shows that the residuals tend to zero when the imputation model fits the observed data better. The distribution of the observed data was more extreme compared with the empirical posterior distributions of replicated data generated under the logistic model with only variable $Z$.

## 6.6 Application

### 6.6.1 Background

We illustrate the application of the proposed PPC for multiple missing variables with the data from the body mass index (BMI) of the Dutch. BMI is defined as the body weight divided by the square of the body height, which is broadly applied to categorize a person into underweight, normal, overweight, and obese. Since measuring a person's weight and height is costly, it is alternative to ask people to report their weight and height. However, such self-report data is systematically biased. People are used to overestimating their height and underestimating their weight, leading to a lower self-report BMI compared with measured BMI [van Buuren, 2018, Section 9.3]. The goal of the study is to estimate unbiased BMI from the self-report data. We apply the multiple imputation approach to fill the unobserved measured weight and height.

The data we analyze is named `selfreport` in `MICE` package. The data consists of two components. One is the calibration dataset contains 1257 Dutch individuals with both self-report and measured height and weight, which taken from Krul et al. [2011]. The origin survey measured 4459 adults in either Italy, Netherlands, or North America aged 18-65 years in 1999 or 2000. The second part is that the survey dataset contains 803 Dutch adults aged 18-75 years with only self-reported data. The survey data were collected in November 2007, either online or using paper-and-pencil methods [van Buuren, 2018, Section 9.3]. Six variables are included in the application: age (years), sex(male or female), hm denoting measured height (cm), hr denoting self-reported height (cm), wm denoting measured weight (kg), and wr denoting self-reported weight (kg).

The aim is to check whether the proposed PPC for MI works for a missing variable when imputation models for other missing variables are not appropriate. Because FCS imputed missing variables in an iteration way, the incorrect imputation model for the missing variable $Y_1$ may disturb the diagnostic for the missing variable $Y_2$. The deviation of the distribution of $Y_1$ may cause a sub-optimal imputation model to become the optimal choice of $Y_2$. We design two linear regression imputation models for hm: one includes all the other variables, and the other includes all the other variables except the variable hr. Similarly, there are two linear regression imputation models for wm: one includes all the other variables, and the other includes all the other variables except the variable wr. In such a case, we have four imputation strategies to evaluate:

1. Case 1: impute hm by linear regression including hr and wm by linear regression including wr.

2. Case 2: impute hm by linear regression including hr and wm by linear regression excluding wr.

3. Case 3: impute hm by linear regression excluding hr and wm by linear regression including wr.

4. Case 4: impute hm by linear regression excluding hr and wm by linear regression excluding wr.

Although it is evident that hr and wr are significant covariates for the imputation of hm and wm, we deliberately ignore these two variables to reflect more discernible discrepancies for the purpose of illustration.

### 6.6.2 Results

The table 6.6 shows that the best imputation models among these four are the one which includes both wr and hr. The average distance and the width of confidence intervals for the observed data was the smallest for both hm and wm. No matter the imputation model of hm was correctly specified, we could conclude that the linear regression imputation model for wm should be based on all the other variables. When fixing the imputation model for the hm (no matter including hr or not), the average distance and the average width of the confidence interval of hm derived under the linear model included hr was remarkably less than the result taken under the linear model excluded the covariate hr. The graphical results (Figure 6.9-6.12) show the same conclusion. When the linear regression imputation model for wm or hm was correctly specified, the imputed data overlapped the observed data in the scatter plot. The observed data would be closer to the center of the confidence interval, and the width of confidence intervals was smaller. However, the result of wr in case 3 was slightly larger than that in case 1. Similarly, the result of wr in case 4 was slightly larger than that in case 2. Two findings imply that incorrect specification of the imputation models for other missing variables $Y_j$ would influence the target variable $Y_j$ on which we perform the PPC. Our application scenario is relatively simple: the linear model is sufficient to reflect the data generating process of missing variables. However, we do not rule out the possibility that under extreme and complicated cases, incorrect specification of the imputation models for other missing variables $Y_j$ would prevent us from selecting the most suitable imputation model for the missing variable $Y_j$. A similar result could be found in fixing the imputation model for the wm (no matter the imputation model includes wr or not). The average distance and the average confidence interval of wm derived under the linear model had wr was remarkably less than the result taken under the linear model excluded the covariate wr.

## 6.7 Discussion

The proposed imputation model diagnostic procedure based on PPC works on continuous and discrete variables under parametric and hot-deck multiple imputation approaches. We could derive more information for a continuous variable, such as the blueprint of imputation model improvement, because of the graphical analysis of the predictive posterior distribution of the observed data. The imputation model selection could be determined with a numerical comparison of summarized statistics, such as the deviation between the observed value and the expectation of corresponding predictive posterior distribution or the width of confidence intervals of predictive posterior distributions for the observed data. The diagnostic could also be performed based on the graphical analysis, such as the scatter plot and the density plot of the observed and replicated data. The more suitable imputation models are, the more similar the replicated data to the observed data in the density and scatter plots. In terms of the distribution plot, the observed data would be closer to the center, and the confidence interval width is narrower. For hot-deck imputation approaches, what PPC diagnoses is the donor selection procedures. However, based on the features of predictive mean matching, the appropriate donor selection does not ensure plausible imputations. Extra analysis of the observed data and the imputed data is necessary. The PPC for categorical data or ordered data is limited since the predictor of the imputation model is the probability of assignment rather than the observed data itself. We currently investigate the residual deviance as the indicator to select the model for categorical data and ordered data.

The application example shows that the PPC works on the multivariate missing datasets. When the imputed covariate deviates from the actual distribution because of the mis-specified imputation models, the imputation model for the predictor could also selected by PPC. In our case study, the misspecification of one missing variable slightly influences the other missing variable's numerical results. However, in more extreme situations, such as a large number of missing variables and more ridiculous imputation models for covariates, the result may be influenced seriously, so as to result in a sub-optimal model selection. Therefore, it is more reasonable to perform the numerical analysis of all missing variables and make the model selection for those variables with remarkably different results under different candidate imputation approaches first.

Existing PPC proposed by He and Zaslavsky [2012] and Gelman et al. [2005] measured the posterior predictive p-value to indicate the discrepancies of summarized statistics between the observed and replicated data. The close to 0 or 1 p-value implies the inadequacy of the imputation model with respect to the target quantities. The target quantities should be calculated with the completed data, which consists of the observed and the imputed data because it allows to calculate the target quantities requiring a complete data matrix. Both He and Zaslavsky [2012] and Nguyen et al. [2015] found that the existing PPC for multiple imputation model is sensitive to the percentage of missing cases. Since the imputed data and the replicated data are generated from the same posterior predictive distribution, with an increasing proportion of missing data, the diagnostic becomes more difficult.

Unlike the existing PPC approach, the PPC discussed in the paper check the imputation model for each missing variable under the FCS framework. We diagnose the distribution of the observed data so that the result would also be reliable with a large proportion of missing cases. The simulation study also shows that the proposed PPC works for different missingness mechanisms. The PPC for multiple imputation models based on target analysis would be more informative when the imputer is also the researcher. The issue would be whether valid inference for the scientific interest could be derived from the imputed data. However, the PPC for multiple imputation models based on the observed data addresses another issue: whether the imputation model is congenial with the substantive model. The imputed data generated after the proposed PPC could be used for more general downstream analysis and different scientific interests.

When the sample size is tremendous, it is better to choose some representative data to check the imputation model so that the scatter plot or the distribution plot would not be too crowded. A clustered procedure could be applied to gather the observed data with closed values and choose one subset in each cluster to check the model. Further investigation is necessary to set the rule to select observed data when the sample size is too large.

|  |  | COV |  |  |  | Average Distance |  |  |  | Average CIW |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | linear model |  | quadratic model |  | linear model |  | quadratic model |  | linear model |  | quadratic model |  |
|  |  | 75%CI | 95%CI | 75%CI | 95%CI | 75%CI | 95%CI | 75%CI | 95%CI | 75%CI | 95%CI | 75%CI | 95%CI |
|  | missingness |  |  |  |  |  |  |  |  |  |  |  |  |
| MCAR | 30 | 0.76 | 0.96 | 0.75 | 0.93 | 2.41 | 2.41 | 0.79 | 0.79 | 6.64 | 11.32 | 2.27 | 3.87 |
|  | 50 | 0.72 | 0.96 | 0.76 | 0.94 | 2.44 | 2.44 | 0.77 | 0.77 | 6.61 | 11.27 | 2.25 | 3.83 |
|  | 80 | 0.77 | 0.95 | 0.78 | 0.94 | 2.25 | 2.25 | 0.87 | 0.87 | 6.53 | 11.13 | 2.51 | 4.28 |
| MARr | 30 | 0.75 | 0.95 | 0.74 | 0.94 | 2.28 | 2.28 | 0.8 | 0.8 | 6.26 | 10.66 | 2.31 | 3.94 |
|  | 50 | 0.76 | 0.95 | 0.73 | 0.95 | 2.21 | 2.21 | 0.81 | 0.81 | 6.3 | 10.73 | 2.3 | 3.91 |
|  | 80 | 0.8 | 0.96 | 0.77 | 0.92 | 1.8 | 1.8 | 0.83 | 0.83 | 5.43 | 9.25 | 2.38 | 4.05 |

Table 6.1: The rate of nominal confidence interval covers the observed data (COV), the mean of the distance between the observed data and the location of the predictive posterior distribution (Distance), and the average width of the nominal (75% and 95%) confidence intervals (CIW) for two imputation models (linear model and quadratic model) under different combinations of experimental factors. The analysis model is a quadratic equation.

|  |  | COV |  |  | Average distance |  |  | Average CIW |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | missingness | PC | SMC-FCS | PMM | PC | SMC-FCS | PMM | PC | SMC-FCS | PMM |
| MCAR | 30 | 0.94 | 0.94 | 0.91 | 0.62 | 0.62 | 0.68 | 3.02 | 3.1 | 3.03 |
|  | 50 | 0.94 | 0.93 | 0.9 | 0.61 | 0.61 | 0.67 | 3.03 | 3.06 | 2.97 |
|  | 80 | 0.96 | 0.94 | 0.91 | 0.61 | 0.64 | 0.65 | 3.06 | 3.14 | 3 |
| MARr | 30 | 0.94 | 0.94 | 0.89 | 0.59 | 0.59 | 0.64 | 2.93 | 2.84 | 2.84 |
|  | 50 | 0.93 | 0.94 | 0.89 | 0.57 | 0.58 | 0.62 | 2.74 | 2.7 | 2.68 |
|  | 80 | 0.93 | 0.97 | 0.93 | 0.56 | 0.56 | 0.61 | 2.61 | 2.64 | 2.69 |

Table 6.2: The rate of nominal confidence interval covers the observed data (COV), the mean of the distance between the observed data and the location of the predictive posterior distribution (Distance), and the average width of the nominal 95% confidence intervals (CIW) for PC, SMC-FCS and PMM under different combinations of experimental factors. The analysis model is a quadratic equation.

|  |  | COV |  |  | Average distance |  |  | Average CIW |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | missingness | PC | SMC-FCS | PMM | PC | SMC-FCS | PMM | PC | SMC-FCS | PMM |
| MCAR | 30 | 0.76 | 0.75 | 0.7 | 0.62 | 0.62 | 0.68 | 1.77 | 1.82 | 1.78 |
|  | 50 | 0.75 | 0.75 | 0.73 | 0.61 | 0.61 | 0.67 | 1.78 | 1.8 | 1.74 |
|  | 80 | 0.78 | 0.76 | 0.75 | 0.61 | 0.64 | 0.65 | 1.8 | 1.84 | 1.76 |
| MARr | 30 | 0.76 | 0.74 | 0.71 | 0.59 | 0.59 | 0.64 | 1.72 | 1.66 | 1.67 |
|  | 50 | 0.74 | 0.72 | 0.69 | 0.57 | 0.58 | 0.62 | 1.61 | 1.58 | 1.57 |
|  | 80 | 0.75 | 0.73 | 0.7 | 0.56 | 0.56 | 0.61 | 1.53 | 1.55 | 1.58 |

Table 6.3: The rate of nominal confidence interval covers the observed data (COV), the mean of the distance between the observed data and the location of the predictive posterior distribution (Distance), and the average width of the nominal 75% confidence intervals (CIW) for PC, SMC-FCS and PMM under different combinations of experimental factors. The analysis model is a quadratic equation.

| | True value | Estimates value | Coverage rate |
|---|---|---|---|
| $\beta_1$ | 1 | 1.008 | 0.934 |
| $\beta_2$ | 1 | 1 | 0.958 |

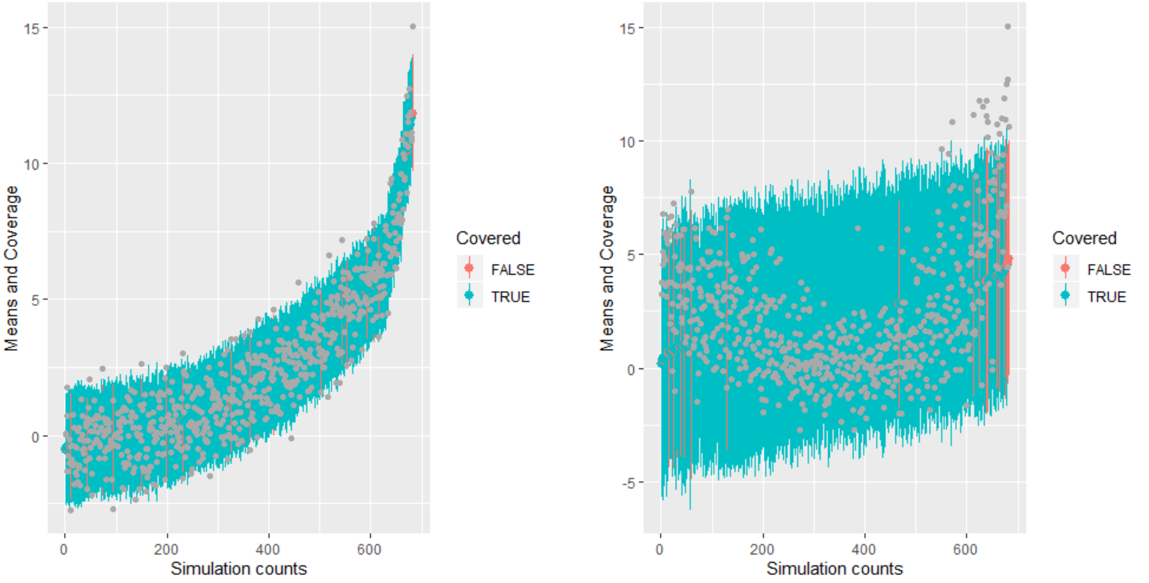
Table 6.4: The PMM performs under the scientific model : $Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon$, where $\alpha = 0$, $\beta_1 = 1$ and $\beta_2 = 1$. The error term and variable $X$ follow standard normal distributions. 30% cases of $X$ and $X^2$ are designed to be jointly missing. The missingness mechanism is MCAR.

| | | mean of residual deviance | |
|---|---|---|---|
| | missingness | with x | without x |
| MCAR | 30 | 0.83 | 1.25 |
| | 50 | 0.85 | 1.27 |
| | 80 | 0.95 | 1.3 |
| MARr | 30 | 0.9 | 1.34 |
| | 50 | 0.94 | 1.35 |
| | 80 | 0.98 | 1.28 |

Table 6.5: The average sum of squared deviance for two imputation models: 1) logistic regression with two preditors $x$ and $z$ 2) logistic regression with one predictor $x$ under different combinations of experimental factors. The outcome is a dichotomous variable $y$ and the binary regression is based on $x$ and $z$.

| | hm | | | wm | | |
|---|---|---|---|---|---|---|
| | cov | average distance | average CIW | cov | average distance | average CIW |
| strategy 1 | 0.95 | 1.57 | 8.27 | 0.95 | 2.28 | 12.46 |
| strategy 2 | 0.95 | 1.65 | 8.89 | 0.94 | 10.9 | 54.38 |
| strategy 3 | 0.95 | 5.58 | 26.89 | 0.94 | 2.35 | 12.83 |
| strategy 4 | 0.95 | 5.56 | 27.88 | 0.97 | 9.84 | 59.57 |

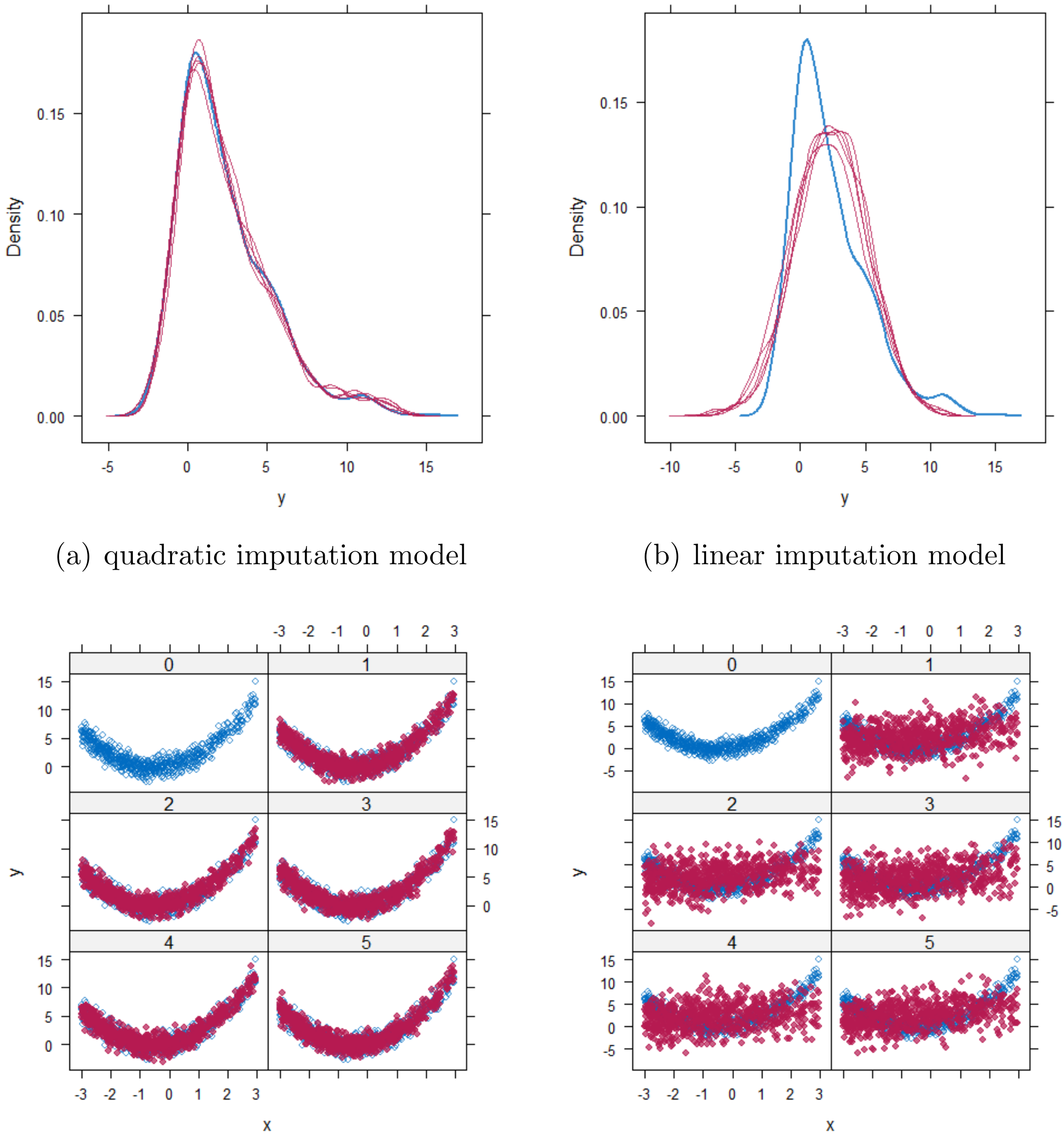
Table 6.6: The performance of 4 imputation strategies summarized by the coverage rate, the average distance and the average width of confidence intervals with respect to missing variables hm and wm



(a) quadratic imputation model

(b) linear imputation model

Figure 6.2: Distribution plots for the first simulation study (quadratic equation with an incomplete outcome) generated under 30% missing cases and MARr missingness mechanism. The confidence interval is 95% nominal.



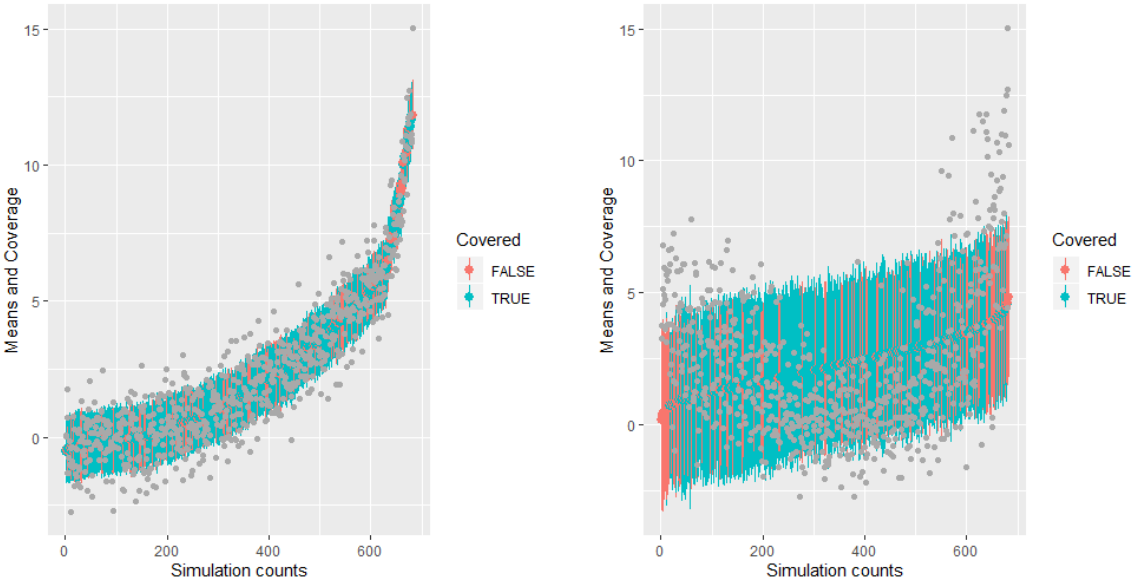
(a) quadratic imputation model

(b) linear imputation model

(c) quadratic imputation model

(d) linear imputation model

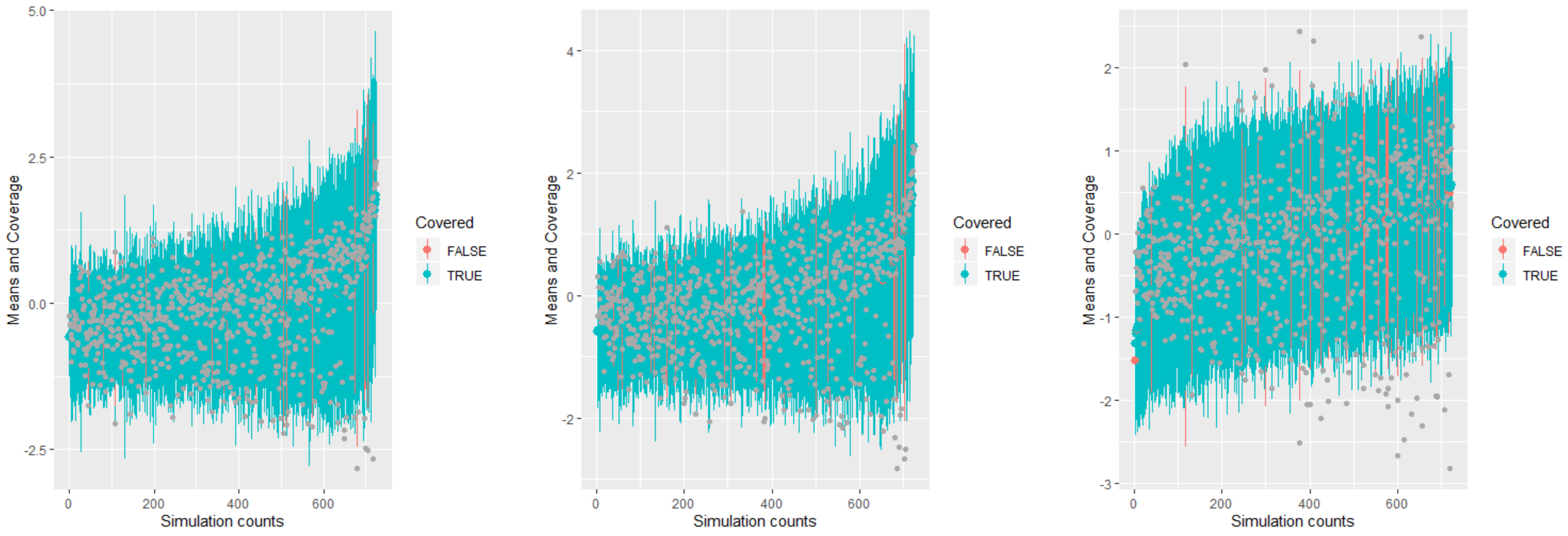
Figure 6.3: Scatterplots and densityplots for the first simulation study (quadratic equation with an incomplete outcome) generated under 30% missing cases and MARr missingness mechanism.



(a) quadratic imputation model

(b) linear imputation model

Figure 6.4: Distribution plots for the first simulation study (quadratic equation with an incomplete outcome) generated under 30% missing cases and MARr missingness mechanism. The confidence interval is 75% nominal.

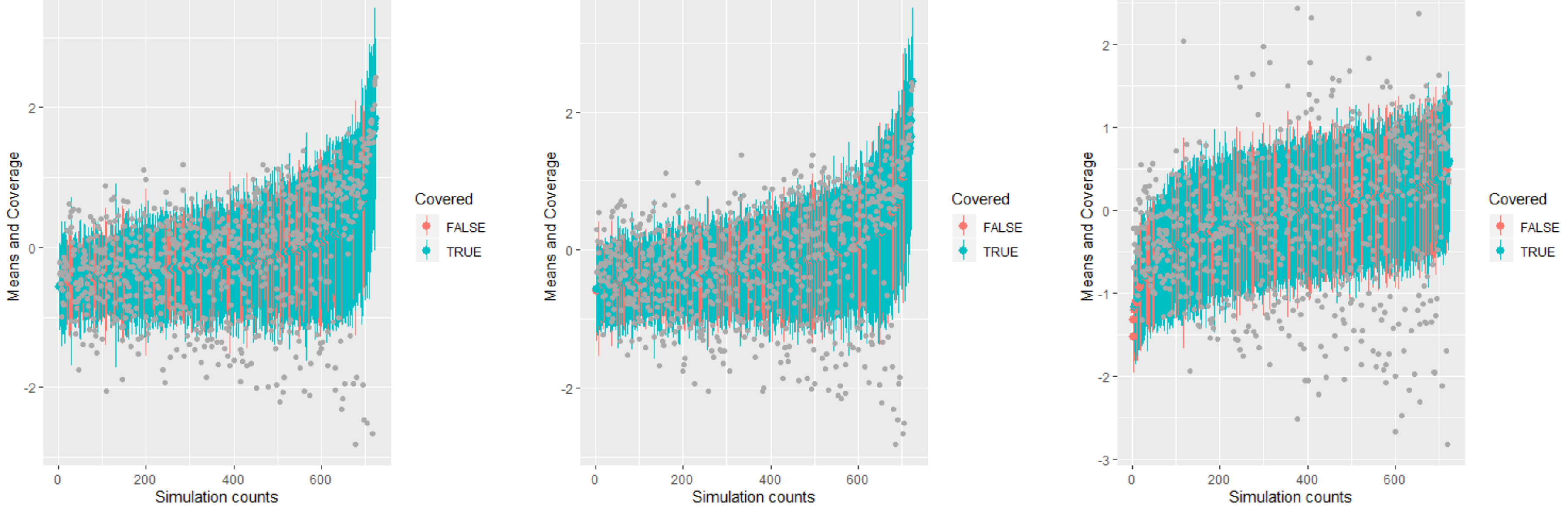


(a) PC

(b) SMC-FCS

(c) PMM

Figure 6.5: Distribution plots for the second simulation study (quadratic equation with incomplete covariates) generated under 30% missing cases and MARr missingness mechanism. The nominal level is 95%.

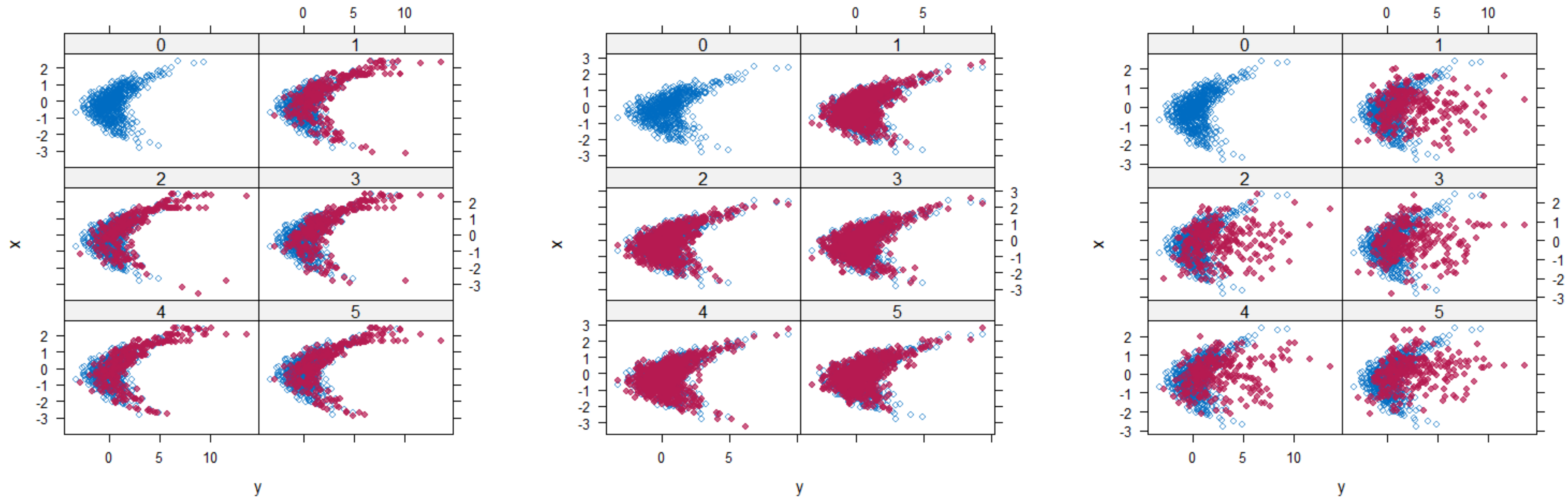


(a) PC

(b) SMC-FCS

(c) PMM

Figure 6.6: Distribution plots for the second simulation study (quadratic equation with incomplete covariates) generated under 30% missing cases and MARr missingness mechanism. The nominal level is 75%.

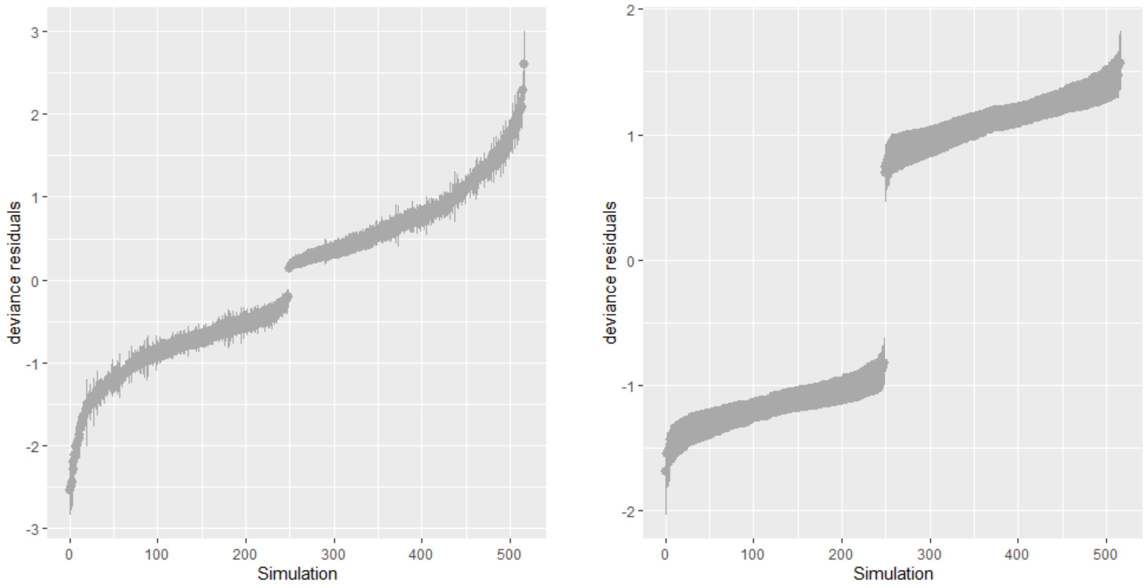


(a) PC

(b) SMC-FCS

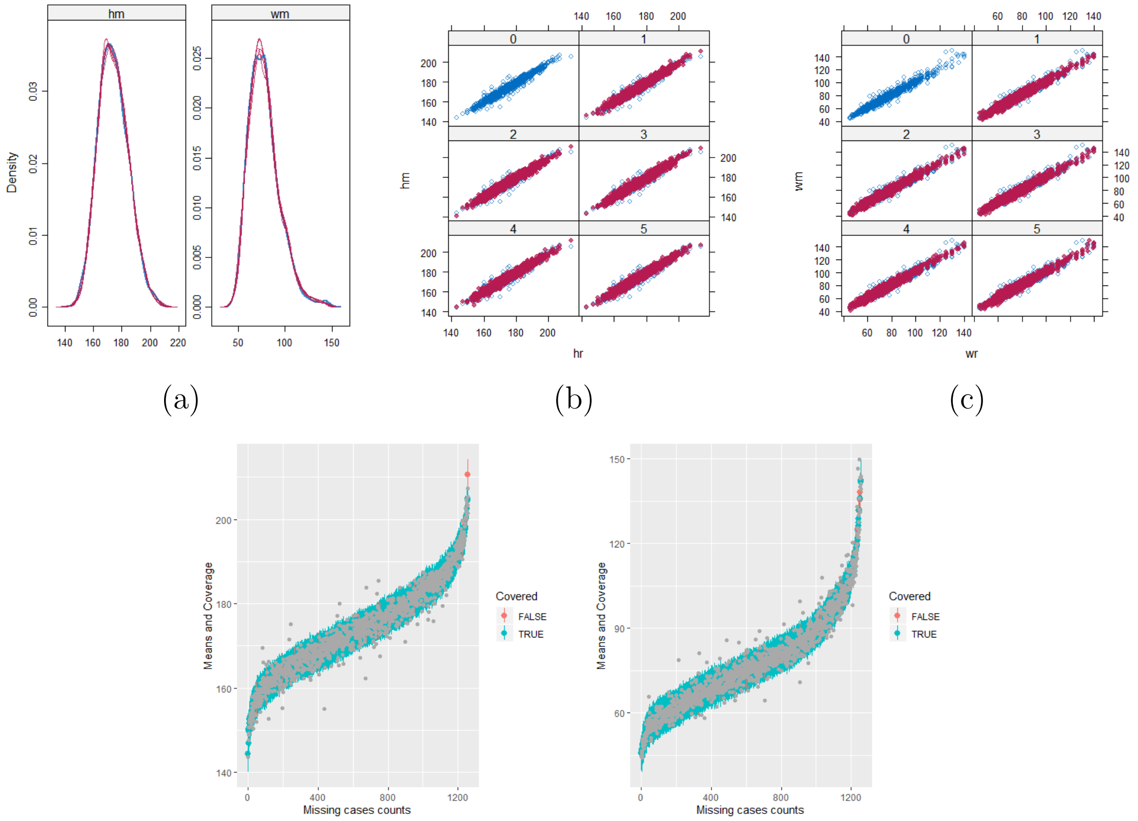
(c) PMM

Figure 6.7: Scatterplots for the second simulation study (quadratic equation with incomplete covariates) generated under 30% missing cases and MARr missingness mechanism.



(a) logistic model based on $X$ and $Z$ (b) logistic model based on $Z$

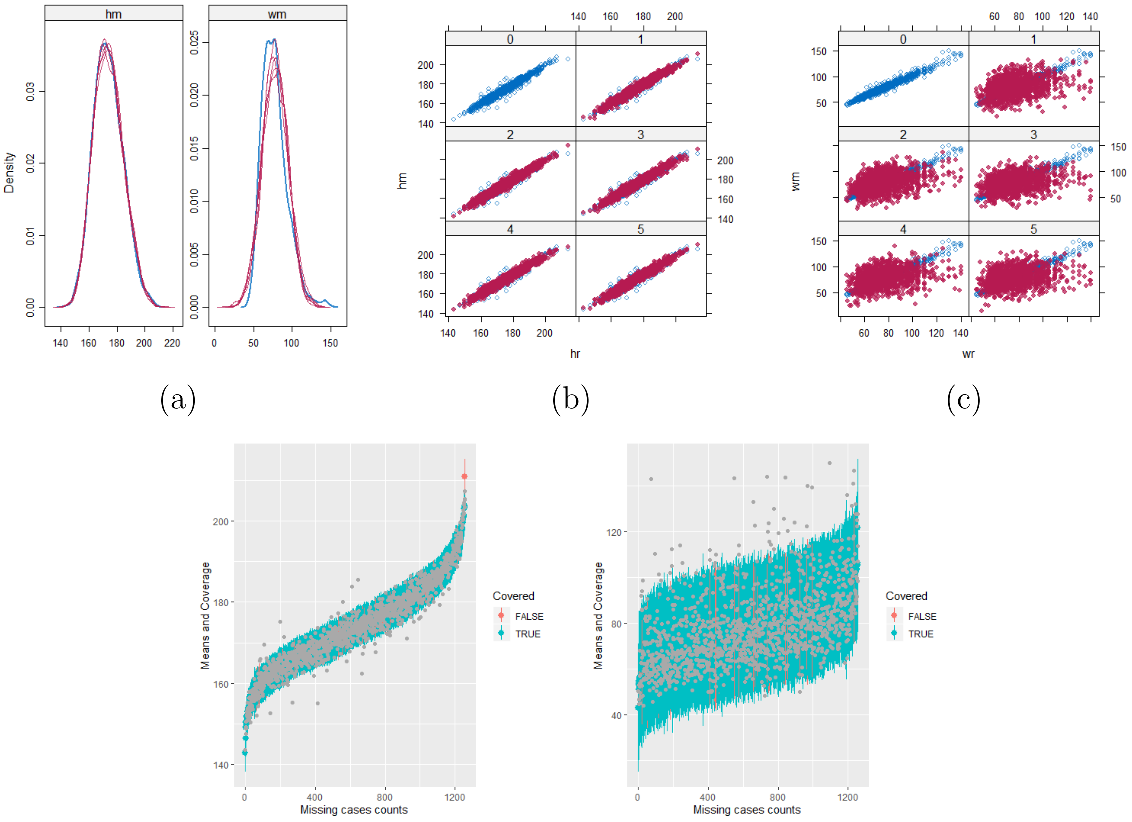
Figure 6.8: The plot of deviance residuals for the third simulation study (generalized linear model for discrete variables) generated under two logistic regression imputation models. The percentage of missing is 30%, and the missingness mechanism is MARr.



(a) (b) (c)

(d) (e)

Figure 6.9: Graphical analysis of the BMI data with imputation strategy case 1



(a) (b) (c)

(d) (e)

Figure 6.10: Graphical analysis of the BMI data with imputation strategy case 2

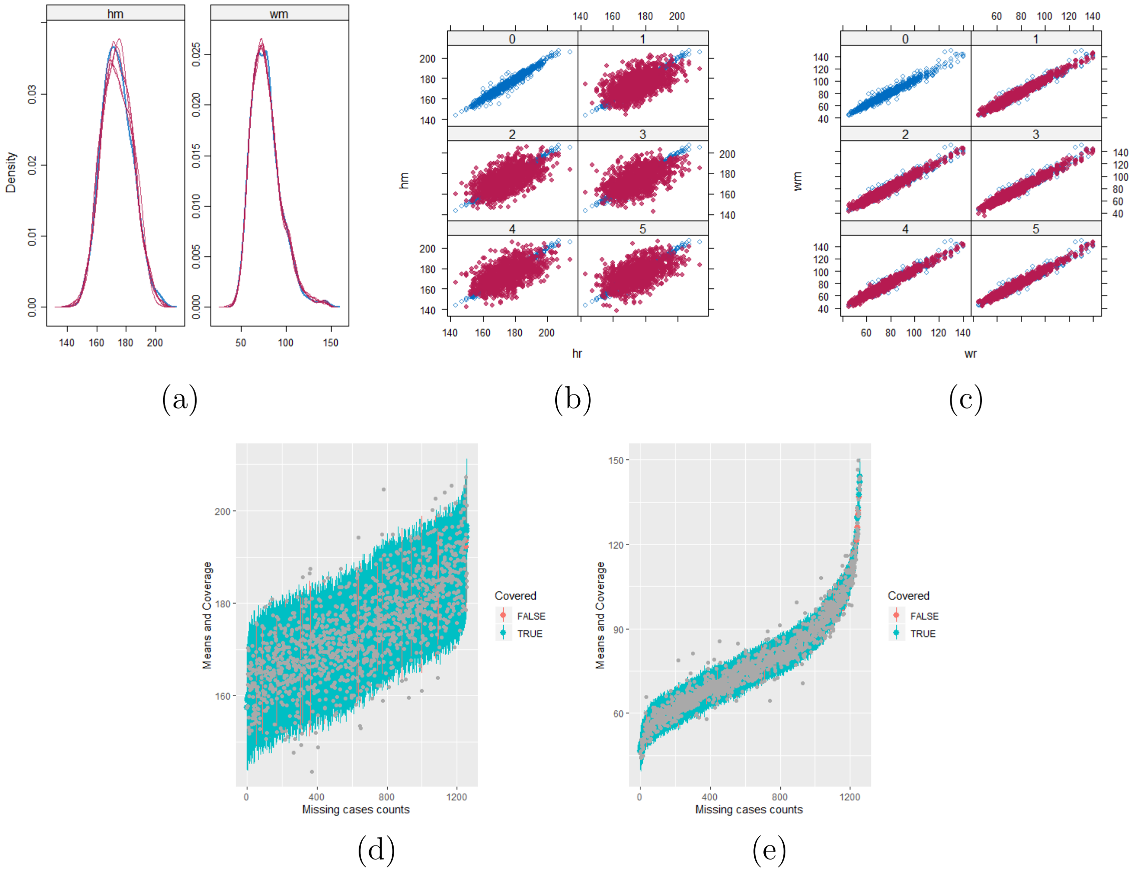
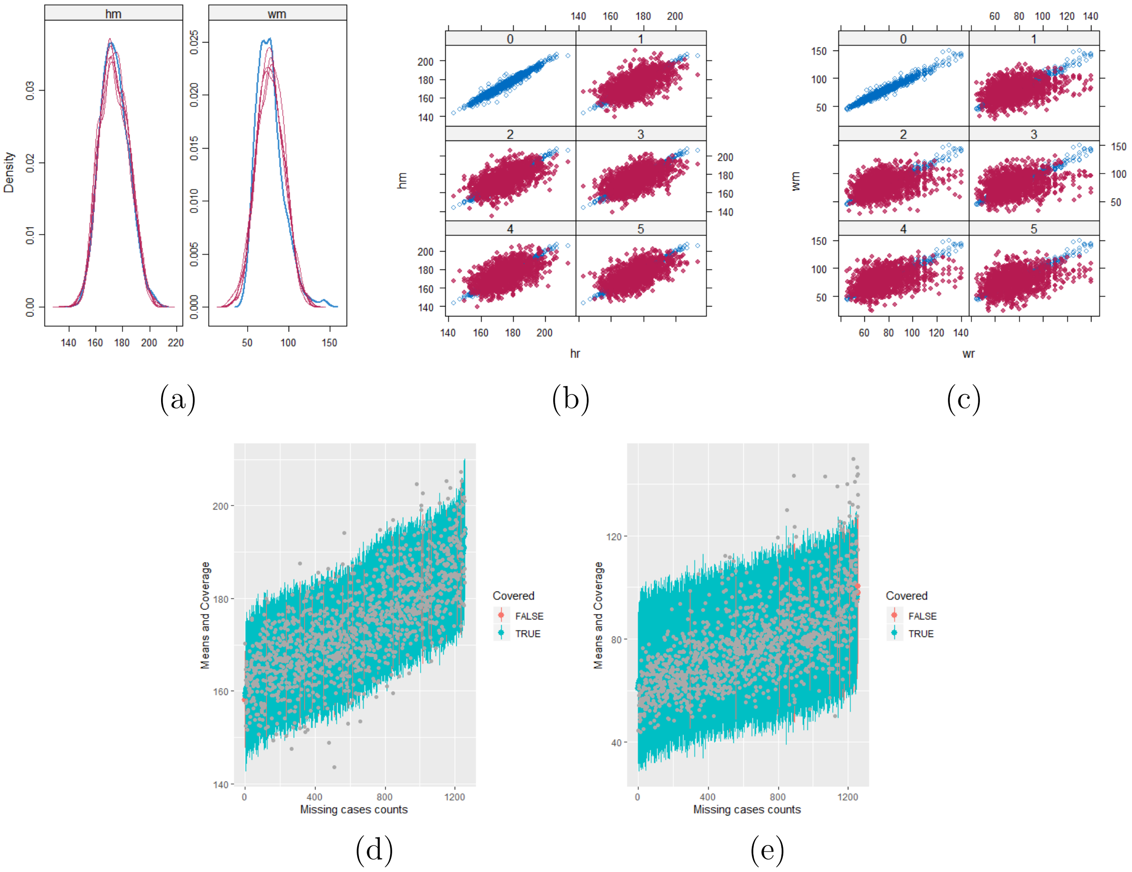


Figure 6.11: Graphical analysis of the BMI data with imputation strategy case 3



(a)

(b)

(c)

(d)

(e)

Figure 6.12: Graphical analysis of the BMI data with imputation strategy case 4

# References

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

J. Albert. *Bayesian computation with R.* Springer, 2009.

P. D. Allison. *Missing data.* Sage publications, 2001.

P. D. Allison. Missing data techniques for structural equation modeling. *Journal of abnormal psychology*, 112(4):545, 2003.

A. Araujo, S. Julious, and S. Senn. Understanding variation in sets of n-of-1 trials. *PLoS One*, 11(12):e0167167, 2016.

B. C. Arnold and S. J. Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Compatibility of partial or complete conditional probability specifications. *Journal of statistical planning and inference*, 123(1):133–159, 2004.

M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

J. W. Bartlett, S. R. Seaman, I. R. White, J. R. Carpenter, and A. D. N. Initiative*. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, 24(4):462–487, 2015.

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

I. Bondarenko and T. Raghunathan. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in medicine*, 35 (17):3007–3020, 2016.

J. Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. 1999.

R. de Jong, S. van Buuren, and M. Spiess. Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics - Simulation and Computation*, 45(3):968–985, jul 2014. 10.1080/03610918.2014.911894.

T. De Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons, 2011.

H. Demirtas. Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods*, 3(2):5, 2004.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (3):655–671, 2016.

P. Ding, F. Li, et al. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018.

P. Ding, A. Feller, and L. Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

H. Djebbari and J. Smith. Heterogeneous impacts in progresa. *Journal of Econometrics*, 145(1-2):64–80, 2008.

M. L. Eaton. Multivariate statistics: A vector space approach. *Lecture Notes-Monograph Series*, 53:i–512, 2007. ISSN 07492170. URL `http://www.jstor.org/stable/20461449`.

R. Finke and A. Adamczyk. Cross-national moral beliefs: The influence of national religious context. *Sociological Quarterly*, 49(4):617–652, 2008.

G. L. Gadbury, H. K. Iyer, and D. B. Allison. Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics*, 11(4):313–333, 2001.

A. Gelman, G. King, and C. Liu. Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443):846–857, 1998.

A. Gelman, I. Van Mechelen, G. Verbeke, D. F. Heitjan, and M. Meulders. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, 61(1):74–85, 2005.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

H. Goldstein, J. R. Carpenter, and W. J. Browne. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(2):553–564, 2014.

J. H. Goodnight. A tutorial on the sweep operator. *The American Statistician*, 33(3): 149–158, 1979.

B. Goodrich, J. Gabry, I. Ali, and S. Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2019. URL `https://mc-stan.org/rstanarm`. R package version 2.19.2.

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

S. Gruber and M. J. Van der Laan. tmle: An r package for targeted maximum likelihood estimation. 2011.

S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335 (15):1081–1090, 1996.

Y. He and A. M. Zaslavsky. Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in medicine*, 31(1):1–18, 2012.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

J. J. Heckman. The scientific model of causality. *Sociological methodology*, 35(1):1–97, 2005.

J. J. Heckman, J. Smith, and N. Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

J. J. Heckman, J. E. Humphries, S. Urzua, and G. Veramendi. The effects of schooling on labor market and health outcomes. *Unpublished Manuscript. University of Chicago, Department of Economics*, 2010.

S. G. Heeringa. Multivariate imputation of coarsened survey data on household wealth. 2001.

D. F. Heitjan and R. J. A. Little. Multiple imputation for the fatal accident reporting system. *Applied Statistics*, 40(1):13, 1991. 10.2307/2347902.

M. A. Hernan and J. M. Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

J. Hoogland, M. van Barreveld, T. P. Debray, J. B. Reitsma, T. E. Verstraelen, M. G. Dijkgraaf, and A. H. Zwinderman. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in medicine*, 39(25):3591–3607, 2020.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

R. A. Hughes, I. R. White, S. R. Seaman, J. R. Carpenter, K. Tilling, and J. A. Sterne. Joint modelling rationale for chained equations. *BMC medical research methodology*, 14(1):28, 2014.

J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

J. G. Ibrahim, H. Chu, and M.-H. Chen. Missing data in clinical studies: issues and methods. *Journal of clinical oncology*, 30(26):3297, 2012.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

A. Z. Israels. *Eigenvalue techniques for qualitative data (M & T series)*. DSWO Press, 1987. ISBN 90-6695-020-X.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

M. K. James R. Carpenter. *Multiple Imputation and Its Application*. John Wiley & Sons, Feb. 2013. ISBN 0470740523.

A. J. Krul, H. A. Daanen, and H. Choi. Self-reported and measured weight, height and body mass index (bmi) in italy, the netherlands and north america. *The European Journal of Public Health*, 21(4):414–419, 2011.

A. Lamont, M. D. Lyons, T. Jaki, E. Stuart, D. J. Feaster, K. Tharmaratnam, D. Oberski, H. Ishwaran, D. K. Wilson, and M. L. Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in medical research*, 27(1):142–157, 2018.

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

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

J. Liu, A. Gelman, J. Hill, Y.-S. Su, and J. Kropko. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014.

N. Longford. Multilevel analysis with messy data. *Statistical Methods in Medical Research*, 10(6):429–444, 2001.

R. P. McDonald. A unified treatment of the weighting problem. *Psychometrika*, 33(3): 351–381, sep 1968. 10.1007/bf02289330.

X.-L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.

C. Moriarity and F. Scheuren. A note on rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 21(1):65–73, 2003.

T. P. Morris, I. R. White, and P. Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1), jun 2014. 10.1186/1471-2288-14-75.

B. A. Mueller, P. Cummings, F. P. Rivara, M. A. Brooks, and R. D. Terasaki. Injuries of the head, face, and neck in relation to ski helmet use. *Epidemiology*, pages 270–276, 2008.

J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

C. D. Nguyen, K. J. Lee, and J. B. Carlin. Posterior predictive checking of multiple imputation models. *Biometrical Journal*, 57(4):676–694, 2015.

C. D. Nguyen, J. B. Carlin, and K. J. Lee. Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, 14(1):8, 2017.

H. I. Oberman, S. van Buuren, and G. Vink. Missing the point: Non-convergence in iterative imputation algorithms. 2020.

A. Olinsky, S. Chen, and L. Harlow. The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, 151(1):53–79, 2003.

T. D. Pigott. Missing predictors in models of effect size. *Evaluation & the Health Professions*, 24(3):277–307, 2001.

R. S. Poulson, G. L. Gadbury, and D. B. Allison. Treatment heterogeneity and individual qualitative interaction. *The American Statistician*, 66(1):16–24, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

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

S. Rässler. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, volume 168. Springer Science & Business Media, 2012.

A. C. Rencher. *Methods of multivariate analysis*, volume 492. John Wiley & Sons, 2003.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

P. Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

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

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

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

J. L. Schafer. *Analysis of incomplete multivariate data.* CRC press, 1997.

J. L. Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8 (1):3–15, 1999.

J. L. Schafer. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica neerlandica*, 57(1):19–35, 2003.

J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

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

R. M. Schouten, P. Lugtig, and G. Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

S. R. Seaman, J. W. Bartlett, and I. R. White. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1):1–13, 2012.

L. Shamseer, M. Sampson, C. Bukutu, C. H. Schmid, J. Nikles, R. Tate, B. C. Johnston, D. Zucker, W. R. Shadish, R. Kravitz, et al. Consort extension for reporting n-of-1 trials (cent) 2015: Explanation and elaboration. *Bmj*, 350:h1793, 2015.

J. Siddique and T. R. Belin. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, 27(1):83–102, 2008.

S. Sinharay, H. S. Stern, and D. Russell. The use of multiple imputation for the analysis of missing data. *Psychological methods*, 6(4):317, 2001.

W. Smink. Towards estimation of individual causal effects: The use of a prior for the correlation between potential outcomes. Master's thesis, 2016.

O. W. Souverein, A. H. Zwinderman, and M. W. Tanck. Multiple imputation of missing genotype data for unrelated individuals. *Annals of human genetics*, 70(3):372–381, 2006.

Y.-S. Su, A. E. Gelman, J. Hill, and M. Yajima. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. 2011.

L. Sun, S. Ji, S. Yu, and J. Ye. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *IJCAI*, volume 9, pages 1230–1235, 2009.

K. Sundell, K. Hansson, C. A. Löfholm, T. Olsson, L.-H. Gustle, and C. Kadesjö. The transportability of multisystemic therapy to sweden: short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology*, 22(4):550, 2008.

C. Tanasoiu and C. Colonescu. Determinants of support for european integration: The case of bulgaria. *European Union Politics*, 9(3):363–377, 2008.

J. Twisk and W. de Vente. Attrition in longitudinal studies: how to deal with missing data. *Journal of clinical epidemiology*, 55(4):329–337, 2002.

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

S. van Buuren. *Flexible Imputation of Missing Data, Second Edition.* Chapman and Hall/CRC, 2018.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations inR. *Journal of Statistical Software*, 45(3), 2011.

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

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

A. L. Van Den Wollenberg. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219, 1977.

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

G. Vink and S. van Buuren. Multiple imputation of squared terms. *Sociological Methods & Research*, 42(4):598–607, 2013.

G. Vink and S. van Buuren. Hybrid imputation, Dec. 2019. URL `https://www.gerkovink.com/London2019/`.

G. Vink, L. E. Frank, J. Pannekoek, and S. van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, jan 2014. 10.1111/stan.12023.

G. Vink, G. Lazendic, and S. van Buuren. Partioned predictive mean matching as a large data multilevel imputation technique. *Psychological Test and Assessment Modeling*, 57 (4):577–594, 2015.

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

D. Westreich, J. K. Edwards, S. R. Cole, R. W. Platt, S. L. Mumford, and E. F. Schisterman. Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology*, 44(5):1731–1737, 2015.

I. R. White, R. Daniel, and P. Royston. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational statistics & data analysis*, 54(10):2267–2275, 2010.

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

L.-M. Yu, A. Burton, and O. Rivero-Arias. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, 16(3):243–258, 2007.

Z. Zhang, C. Wang, L. Nie, and G. Soon. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5):687–704, 2013.

J. Zhu and T. E. Raghunathan. Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511):1112–1124, 2015.

# List of Figures

# List of Tables