Distribution Agreement

Signature:

____________________________ ________________

Domonique Watson Hodge Date

Robust Statistical Methods for Handling Missing Data

By

Domonique Watson Hodge

Doctor of Philosophy

Biostatistics

Qi Long, Ph.D.
Advisor

Michael Frankel, M.D.
Committee Member

Yi-An Ko, Ph.D.
Committee Member

Robert Lyles, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Robust Statistical Methods for Handling Missing Data

By

Domonique Watson Hodge
M.S., Florida State University, 2012
B.S., Columbus State University, 2009

Advisor: Qi Long, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2016

Abstract

Robust Statistical Methods for Handling Missing Data

By

Domonique Watson Hodge

Since the dawn of data collection, researchers have faced the problem of missing data. There are a multitude of reasons data may be missing. Appropriately dealing with missing data requires a careful examination of the data to identify the source, pattern, and missing data mechanism. It is well known that a naive analysis without adequate handling of missing data reduces statistical power, results in loss of efficiency, and potentially biases parameter estimates which can ultimately lead to invalid conclusions. Multiple imputation (MI) is one of the most widely used methods for handling missing data. The key idea of MI is to replace each missing value with a set of plausible values drawn from their predictive distributions conditional on the observed data. Multiple imputed data sets are generated to account for uncertainty of imputing missing values. We review the terminology and current literature on missing data in Chapter 1.

In Chapter 2, we aim to develop MI methods to handle missing data in the presence of high-dimensional data where the missing data mechanism is assumed to be ignorable. Existing (MI) methods implemented in most statistical software are not applicable or do not perform well in the high-dimensional setting where the number of predictor is large relative to the sample size. To remedy this issue, we develop an MI approach that uses dimension reduction techniques. Specifically, our approach uses sure independent screening (SIS) followed by either sparse principal component analysis (sPCA) or sufficient dimension reduction regression in constructing imputation models in the presence of high-dimensional data. Our extensive simulation studies demonstrate that in the presence of high-dimensional data using SIS followed by

sPCA to perform MI achieves better performance than the other imputation methods including several existing imputation approaches. We further illustrate our approach using gene expression data from a prostate cancer study.

In Chapter 3, we develop nonparametric imputation methods to handle non-ignorable missing data. Most imputation techniques are designed for ignorable missing data since non-ignorability is an assumption more challenging to handle. Under non-ignorable missingness, one assumes the nonresponse mechanism depends on unobserved values, and the outcome model for the variable with missing values and the nonresponse model must be modeled jointly. Consequently, joint modeling can produce results that are sensitive to the misspecification of the outcome and nonresponse models. We propose a nonparametric method for handling non-ignorable missingness via bootstrap imputation and multiple imputation. The key idea underlying our proposed approach is to formulate two working models for the outcome and for nonresponse, respectively. Using the two working models, we derive predictive scores which achieves dimension reduction and use the resulting scores coupled with a nearest neighbor hot deck to multiply impute missing values. Our approach allows users to incorporate prior knowledge on the working models through the use of weights. Compared with the existing MI methods, our approach is more robust to misspecification of the two models and allows for a natural sensitivity analysis. The proposed bootstrap imputation approach is shown to outperform several existing multiple imputation methods for non-ignorable missing data in simulations. In addition, the method is illustrated using data from the Georgia Coverdall Acute Stroke Registry.

In Chapter 4, we aim to evaluate diagnostic methods for imputation models assuming a non-ignorable missing data mechanism. Most of the existing diagnostic approaches have been developed assuming the missing data mechanism is ignorable. As a consequence, they are not directly applicable to our nonparametric imputation methods for nonignorable imputation methods. To address this issue, we adapt posterior predictive checking with the posterior predictive p-value as the summary measure to assess the performance of imputation models under non-ignorable missingness. In simulations, we correctly and incorrectly specify the imputation models and determine whether posterior predictive checking is useful in detecting discrepancies in the misspecified imputation model. Our extensive simulations suggest that, in the settings we evaluated, posterior predictive p-values can be useful in diagnosing deficiencies in non-ignorable imputation models. We also illustrate this approach using the Georgia Coverdall Acute Stroke Registry. In Chapter 5, we present potential future work to extend our methodologies to handle additional problems that arise from missing data.

Robust Statistical Methods for Handling Missing Data

By

Domonique Watson Hodge
M.S., Florida State University, 2012
B.S., Columbus State University, 2012

Advisor: Qi Long, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2016

Acknowledgement

I would like to express the deepest appreciation to my advisor Dr. Qi Long, who always encourages me to work harder and is patient. He continually conveyed a spirit of adventure in regard to research, and an excitement in teaching. He really got me passionate about conducting research in biostatistics. After taking his class and learning more about missing data, I thought it would be a great research topic that was applicable to almost any research study. Although at times research seemed impossible to complete, I was always eager to do more. Without Dr. Long's guidance and persistent help, I would not be able to complete my doctoral studies. For this I am forever grateful. In addition, I would like to thank my committee members, Dr. Lyles, Dr. Ko, and Dr. Frankel for their support and dedication in regards to helping guide my research.

I would also love to thank my husband, Samuel Hodge, who always was supportive of me continuing my education. He solely supported our family as I continued my studies, and for this I owe him my life. I love and appreciate him more than he could ever imagine. I also really appreciate the support of my mom. She raised my brothers and I as a single parent. I witnessed her struggles, but she always tried to hide them so that nothing would distract us in school. My mom always encouraged me in school and told me that I could do anything that I put my mind to. My mother was never able to get a college degree, but I promised her that I would obtain the highest degree possible. She is my true inspiration and the true reason I made it this far. I certainly have to mention my in-laws who were very supportive in helping me raise my favorite person in the world, my son Samuel Hodge II. They also continuously encouraged me and always went above and beyond to help make graduate school life a little easier for me.

I really value the mentorship of Dr. Felicia Williams, Dr. Renee Moore, Dr. Donna Watts, and Dr. Amita Manatunga. These women were not only role models, but were instrumental in me developing my skills as a doctoral student. Most importantly, they were always there when I needed a shoulder to cry on or wanted to vent about life.

Last but not least, I would like to thank our hard working department chair Dr. Waller, all my professors, and staff in the Biostatistics department. Also, special thanks to Professor Andrew Hill who helped me form better study habits and tutor me so that I could successfully pass my qualifying exam. He was always patient and would explain questions hundreds of ways until he knew I completely understood. Of course, I would also not be here without his guidance. I would also like to thank the Initiative to Maximize Student Diversity, Patient Centered Outcome Research Institute, and SAS Advanced Analytics Division for their financial support granted through predoctoral fellowship.

Contents

List of Tables

Chapter 1

Introduction

1.1 Missing Data Problem

Missing data are a common problem in data collection. There are many potential sources of missing data. For instance, in surveys information may be missing due to nonresponse from subjects because participants may refuse to reveal some private information or participants may forget to answer certain questions. Some survey developers may design the study so that some questions are asked of only a subset of participants. Longitudinal studies may suffer from missing data due to attrition. Subjects may drop out before the end of the study because they have moved away from the study location, see no personal benefit in participating, or become decease. In registry data, clinicians may fail to collect some data on the patient or some questions may not be applicable to the patient and are skipped. Appropriately dealing with missing data requires a careful examination of the data to identify the source, pattern, structure, and missing data mechanism. Understanding these aspects are vital since the different methods for dealing with missing data make assumptions about the missingness.

1.2 Missing Data Nomenclature

We can distinguish between patterns of missingness. Some methods to handle missing data apply to any pattern of missing data while some are restricted to special missing data patterns. Little and Rubin (2014) describe the missing-data patterns as univariate, monotone, file matching, latent variable, general, unit nonresponse, and item nonresponse. Univariate nonresponse occurs when missingness is confined to a single variable. In practice, missing values often occur in multiple variables. A monotone missing data pattern is typically associated with a longitudinal study where participants drop out. The data have a monotone missing pattern when the event of a particular individual implies that all subsequent variables are also missing.

The file-matching problem exists when data are combined from two different sources that capture different data. Latent-variable patterns exist when unobserved, latent variables are completely missing. Perhaps the most common configuration of missing values is a general missing data pattern. A general pattern has missing values randomly dispersed throughout the data. Based on this typology, determinants of missing data can also be distinguished. For example, unit nonresponse can occur when a questionnaire is administered and a subset of sampled individuals do not complete the questionnaire. Unit nonresponse is often divided into three components to include non-contact, inability to respond, or refusal. Alternatively, item nonresponse occurs when the sample unit does not respond to some items or questions which ultimately cause missing data. Understanding the missing data patterns is important because inadequate handling of missing data may lead to biased estimation and inference.

To understand the mechanisms that lead to missing data, Rubin (1976) created a taxonomy for missing data which is widely used in the literature and determine the appropriateness of methods to handle its complexities. The missing data mechanism can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). The most basic assumption about missing data is to assume the data are MCAR. MCAR is an assumption in which the missingness of the data is completely independent of both the observed and the missing values. To introduce these terms, suppose we have complete data $Y = (y_{ij})$ where $i = 1, .., n$ defines the number of subjects, and $j = 1, ..., p$ defines the p variables. Suppose Y is subjected to missingness where Y_{obs} is the observed components of Y and Y_{mis} are the missing components. We let $R = R_{ij}$ be the response matrix which takes values 1 if y_{ij} is observed and 0 if y_{ij} is missing. In terms of the conditional distribution of R given Y, if $P(R|Y) = P(R)$, then the data are MCAR. MAR is a less restrictive assumption that implies that missingness does not depend on missing components of

Y but does depend on observed components. That is, $P(R|Y) = P(R|Y_{obs})$. If the data is neither MCAR or MAR then the probability that Y is missing depends on unobserved values and is missing not at random (MNAR). MNAR informs that the distribution of the observed values differs from the distribution of the missing values. From the perspective of surveys, the responders differ from the nonrespondents. The latter is a more complex situation to handle because missingness actually depends on values that are not observed. An example to further demonstrate these mechanisms is by the modeling of weight, Y, as a function of a fully observed variable gender. Some subjects may have no record of weight and it is important to understand which mechanism caused the data to be missing. If the data are MCAR, then there is no particular reason why some subjects revealed their weights and others did not. In the case of MAR, one gender may be less likely to disclose its weight, that is, the probability that Y is missing depends on a variable that is completely observed. In the case of MNAR, heavy or light people may be less likely to disclose their weight. That is, the probability that Y is missing depends on the unobserved value of Y itself. The mechanism is important to consider in choosing a missing data approach.

The missing data mechanism can also be referred to as ignorable or nonignorable (Rubin, 1987). The missing data mechanism is said to be ignorable if the data are MCAR or MAR and the parameters governing the missing data process are distinct from parameters to be estimated. Furthermore, if the missing data mechanism is ignorable, then there is no need to postulate a model for the missing data mechanism. A more complicated situation arises when the missing data mechanism is nonignorable. Nonignorability occurs when the data mechanism is either MNAR or MAR and there is no distinction from the parameters. Thus nonignorability requires one to jointly model the data missing data and the response indicator R. Effective estimation in the presence of nonignorable missing data requires some prior knowledge about the missing data mechanism. Although there is a distinction between MNAR and nonignorability, it is often treated as synonymous in practice. It is also important to note that the terms MNAR and NMAR (not missing at random) are often used interchangeably, therefore we adopt this terminology throughout this dissertation.

1.3 Methods to Handle Missing Data

The missing data mechanisms are the basis for understanding the appropriate missing data methods. In this section we present methods to handle missing data by the three missing data mechanisms.

1.3.1 Missing Completely at Random Methods

In the case of MCAR data, the subjects in the sample with completely observed data can be viewed as a random sample of the population of interest. Thus, the commonly used complete-case (CC) analysis method, which discards subjects with missing observations, is appropriate and leads to unbiased results. However, there is a loss of efficiency and decrease in power due to decreased sample size. Another major drawback of the CC method is that often too many observations are discarded and there is a substantial loss in efficiency. An alternative approach is available-case analysis. In an available-case analysis, all cases where the variable of interest is present are included. Available-case analysis uses all possible information in each analysis. While available-case analysis retains more of the observed data as compared to CC, a disadvantage of available-case analysis is that the sample base changes from model to model and will complicate model selection. Another approach, used primarily in longitudinal studies, is last observation carried forward (LOCF). For each subject, missing values in a variable are replaced by the last observed value of the variable. However, LOCF relies strongly on the assumption that the variable value remains unchanged after drop-out. It replaces the missing values with a single imputed value which does not account for the uncertainty of imputation. When the data are not MCAR, then the data can not be viewed as a random sample of the population, therefore, the CC, available-case, and LOCF methods are typically not appropriate. For this reason, Little (1988b) developed a test to determine whether the data are MCAR, or not, which uses a global test statistic that utilizes all of the available data.

1.3.2 Missing at Random Methods

MAR is a less restrictive assumption and is often more suitable. Most inference with missing data have been developed with the assumption that the missingness is at random. CC analysis can also be unbiased under certain MAR mechanisms. The key is that misssingness is conditionally independent of the outcome Y(White and Carlin, 2010). Nevertheless, CC methods generally are not applicable under the assumption of a MAR mechanism. It is also important to understand the dimension of the MAR data because this also determines appropriateness of methods to handle missing data. When data are high-dimensional, for example in gene expression data, the number of p variables is relatively large or strictly larger than the n samples ($p > n$ or $p >> n$, respectively) and traditional missing data methods are not applicable.

1.3.2.1 Techniques for low-dimensional data with missing values assuming MAR

Schafer and Graham (2002) reviewed missing data methods assuming the data are MAR and the sample size larger than the number of predictors. The authors review maximum likelihood approaches, expectation-maximization (EM) algorithm, weighting methods, and multiple imputation (MI). Under MAR, likelihood approaches are commonly used. In the case of an ignorable missing data mechanism, we treat R as a random variable and specify the full distribution of Y and R where β is the parameter for the distribution of Y. The β are distinct from the unknown parameter α in the distribution of the missing data mechanism. Furthermore, likelihood-based inferences for β in the full likelihood are the same as the ignorable likelihood

$L(\beta|Y_{obs}) = \int f(Y_{obs}, Y_{mis}|\beta)dY_{mis}$ which ignores the missing-data mechanism, since the likelihoods are proportional. Kenward and Molenberghs (1998) determined that standard errors and confidence intervals under MAR using the likelihood approach should be based on the observed information matrix instead of the expected information matrix. If there is no closed form expression for the ML estimates, then the EM algorithm (Dempster et al., 1977) is generally used. The EM algorithm is a very general iterative algorithm for maximum likelihood estimation in missing data problems. The E step of the EM algorithm finds the conditional expectation of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for the missing data. The M step of EM determines the estimated parameters by maximizing the expected complete-data loglikehood.

Apart from ML and EM, inverse probability weighting (IPW) and MI can be used when the data are MAR. IPW reconstructs the full population by reweighting the data from subjects who have observed data and deriving these weights from a model for the probability of missingness. Scharfstein et al. (1999) improve the efficiency of the IPW estimator and create a doubly robust augmented IPW estimator. An estimate is doubly robust if it remains consistent when either a model for predicting the missing probabilities or a model for predicting the missing values is correctly specified. An alternative approach is imputation. It is applicable to any type of data and model. It allows for complete-data methods (methods that would be used in the absence of missing data) to be used independent of the imputation. Single imputation, for instance, replaces the missing values with means or draws from the predictive distribution of the missing values. However, single imputation treats the single value as known which does not fully reflect sampling variability and does not account for the uncertainty in nonresponse, therefore, MI is a more favorable approach. MI is a

Bayesian approach which 'fills in' the missing values with M plausible values (Rubin, 1978) drawn from the posterior predictive distribution of Y_{mis}. Furthermore, if we assume the data are MAR, that is, $f(R|Y,\alpha) = f(R|Y_{obs},\alpha)$ where α denotes the set of parameters associated with the missing data mechanism and we also assume that the missing data mechanism is ignorable (ignorability implies α and β are distinct) then it follows that

$$f(Y_{mis}|Y_{obs}) \quad = \quad \int f(Y_{mis}|Y_{obs},\beta) f(\beta|Y_{obs}) d\beta \qquad (1.1)$$

where $f(\beta|Y_{obs})$ is the observed-data posterior distribution and $f(Y_{mis}|Y_{obs},\beta)$ is the conditional predictive distribution (Schafer, 1999). An imputation for Y_{mis} can be created by first randomly drawing the unknown parameters from the observed-data posterior and then proceeding with a random draw of the missing values from the conditional posterior predictive distribution. In order to conduct MI, one must postulate a statistical model for the conditional predictive distribution and the observed-data posterior distribution. A total of M random draws from the conditional predictive distribution create M imputations for the missing values resulting in M completed data sets. Finally, standard complete-data methods are applied to each M data set and the resulting estimates are combined to create one inference which can properly reflect the uncertainty due to nonresponse and validly reflect sampling variability.

1.3.2.2 Techniques for high-dimensional data with missing values assuming MAR

In the case of high-dimensional data ($p > n$) that is assumed to be MAR, existing methods do not perform well. Procedures to handle missing data in the presence of a MAR mechanism with high-dimensional data typically involve imputation. However, the statistical model for the observed-data posterior distribution needs to be adjusted to handle the structure and size of the data. Previous approaches use model trimming or regularization, coupled with imputation. Stekhoven and Bühlmann (2012) used a classification technique, namely random forest (RF), to impute missing values in high-dimensional data. The variable with missing values is treated as the response variable and other (auxiliary) variables are used for bootstrap aggregation of multiple regression trees to potentially reduce overfitting. The predictions are combined from trees to improve accuracy of prediction of the missing values. However, the selection of tuning parameters such as the number of trees and number of nodes needs further investigation. Liao et al. (2014) used a K-nearest-neighbor imputation to fill in missing values. For a missing value, the method seeks its K nearest variables or subjects and imputes by a weighted average of observed values of the similar neighbors. Although the method was shown to perform well in their simulations where the performance was evaluated based on comparisons between true and imputed values, it does not properly propagate the uncertainty in estimating the parameters in the imputation model and hence it is not proper in the sense of Rubin (1987). Improper imputation can lead to biased inference in the subsequent analyses.

Zhao and Long (2013a) proposed an MI approach for high-dimensional data based on regularized regression that does account for the uncertainty in imputation. Specifically, they investigated the use of MI through direct and indirect use of regularized regression. In the former, regularized regression is used for both variable selection and parameter estimation for imputation models; in the latter regularized regression is only used for model trimming. Direct use of regularized regression in MI was shown to achieve superior performance in the settings considered in their work. They also proposed an MI method using the Bayesian lasso (Park and Casella, 2008) to estimate and select important variables in imputation models. However, these methods also have some limitations and particularly they may not yield good performance when the true imputation model is large. There only exist a few methods to handle missing data in high-dimensions; all of which range in ability to reduce bias, efficiently estimate standard errors, and computational efficiency. There is a need to develop new methods to handle missingness in high dimensions. This dissertation proposes a new method to impute missing data by screening for important variables and using linear combinations constructed by either principal components or sufficient dimension reduction regression to build imputation models for multiple imputation. Our methods do not rely on numerous tuning parameters and properly account for the uncertainty across imputations.

1.3.3 Missing Not at Random Methods

It is often really difficult to distinguish between a MAR or MNAR missing data mechanism and no formal test have been developed to distinguish between the two mechanisms. If questions arise about the missing data mechanism, then the MNAR assumption may be more appropriate. However, ignoring the missing data mechanism in the case of MNAR data can result in biased inferences. To remedy this issue, several approaches to handle MNAR data have been developed and include the use of selection models, pattern-mixture models, hot deck imputation, and multiple imputation. To introduce these methods, suppose we have the partition $Y = (Y_{obs}, Y_{mis})$ and X which represents a set of covariates that are fully observed. Heckman (1979) proposed selection models which specify the full-data distribution and require assumptions for the distribution of the missing values. In selection models, the joint distribution of Y and the response indicator R are factorized as $P(Y, R|X) = P(Y|X)P(R|X, Y)$, where the first factor characterizes the distribution of Y and the second distribution models the missing data as a function of Y. Another approach to handle MNAR data uses the methods of Little (1993) who introduced the pattern-mixture models which stratify the responses by missing data patterns. That is, the nonresponse process is a mixture model of varying missing data patterns. The pattern-mixture model are factorized as $P(Y, R|X) = P(Y|X, R)P(R|X)$, where the first distribution characterizes the distribution of Y in the strata defined by different patterns of missing data and the second distribution of R models the incidence of the different patterns. An important practical problem with pattern-mixture models is that the patterns with missing data typically have one or more inestimable parameters, as a result, pattern-mixture models are often underidentified (Little, 1993). Moreover, both approaches are sensitive to model specification, therefore Little and Rubin (2014) recommend sensitivity analysis be performed by estimating a variety of missing data models rather than to rely exclusively on one model.

1.3.3.1 Parametric Imputation Techniques for MNAR data

Imputation-based alternatives to model-based methods allow for standard complete-data methods (methods that would be used in the absence of missing data) and are often preferred due to its ease of implementation. Imputation methods can be either parametric or nonparametric. Glynn et al. (1993) were the first to contribute to the development of parametric multiple imputation in the context of nonignorable missing data. Their approach is applicable when direct information is available from a complete sample of the nonrespondents which is the ideal situation. They considered a mixture model which assumes the distribution differs for respondents and nonrespondents. The authors described the use of multiple imputation in the estimation of the mean and regression parameters when follow-up data is available on the nonrespondents. Their method makes use of three types of subjects: respondents, respondents that were once nonrespondents until follow-up data was obtained, and nonrespondents. The standard approach to estimate the mean is the double-sampling procedure which is the weighted mean of the original respondents and the nonrespondents available for follow-up. The weights are equivalent to the proportion of respondents and the number of nonrespondents before follow-up data is obtained.

The authors discussed alternatives to the double-sampling procedure that are based on mixture models assuming a normal distribution with multiple imputation which are more easily implemented and can handle more complex estimation problems. The normal imputation method was conducted by drawing samples for the nonrespondents from a normal distribution with mean and variance equivalent to the mean of the followed-up nonrespondents. They showed that asymptotically, the multiple imputation estimate and the double-sampling estimate of the mean and standard error are equivalent. The method is shown to perform well in simulation settings for the mean with varying percents of missing values and different nonresponse models. Glynn et al. (1993) believed their method is more appealing to users because the mixture model approach does not involve specification of the model for the probability of nonresponse as compared to selection model based approaches which involve joint specification of the data model and response model. Nevertheless, in practice it can be really difficult to obtain complete follow-up information from a random sample of nonrespondents and selection model based approaches may be favored.

Other methods for multiple imputation with nonignorably missing data include Carpenter et al. (2007) who also proposed a parametric approach which involves reweighting to investigate the influence of departures from the ignorable (MAR) assumption on parameter estimates. The authors presented a simplified clinical trial data setup supposing the data is composed of n subjects, a single response Y_i, baseline data X_i, and a response indicator R_i denoted 1 if Y_i is observed and 0 if Y_i is missing for a subject. Then the nonresponse model is $\text{logit}[P(R_i = 1)] = \alpha + \beta I[\text{patient } i \text{ on active treatment}] + \gamma X_i + \delta Y_i$ where I is an indicator function, α is the adjusted log-odds of observing Y_i, β is the adjusted change in the log-odds ratio of observing Y_i if the subject is assigned to the active treatment, γ is the adjusted change in the log-odds ratio of observing Y_i for a one-unit change in X_i, and δ is the change in the log-odds ratio of observing Y_i for a one-unit change in the single response Y_i. The value of δ determines how much missingness depends on Y and if we suppose $\delta = 0$, then the missing data mechanism is MAR. The δ can be varied for sensitivity analysis.

To avoid the task of jointly fitting a model for the data model and the nonresponse model, a necessity for imputations under nonignorability, Carpenter et al. (2007) derive weights applied to the estimates obtained from the MAR imputations. Importance sampling is used to obtain the weights since one can readily sample from the MAR model as long as the NMAR models estimates are on the same support. A thorough derivation of weights are given using the importance sampling technique with the weight defined as

$$\begin{aligned} \tilde{w}_m &= \exp(\sum_{i=1}^{n_2} -\delta Y_i^m) \\ w_m &= \frac{\tilde{w}_m}{\sum_{m=1}^{M} \tilde{w}_m} \end{aligned}$$

where n_2 is the number of nonrespondents and δ can be specified depending on the assumption of MNAR. The MNAR estimator and its associated variance can be computed as

$$\begin{aligned} \hat{\theta}_{MNAR} &= \sum_{m=1}^{M} w_m \hat{\theta}_M \\ V_{MNAR} &\approx \tilde{V}_W + \left(1 + \frac{1}{M}\right) \tilde{V}_B \end{aligned}$$

where $\hat{\theta}_M$ is the complete data estimates obtain from the MAR imputations, $\tilde{V}_W = \sum_{m=1}^{M} w_m \hat{\sigma}_m^2$ and $\tilde{V}_B = \sum_{m=1}^{M} w_m (\hat{\theta}_m - \hat{\theta}_{MNAR})^2$. The weighting approach is fairly simple and they argue that it is appealing to analyst because of its ease of implementation since there is no need to jointly model the data model and the nonresponse model. Their method requires $M \geq 50$ imputations because accuracy of the approximation of the MNAR estimate and its corresponding variance improves with increasing number of imputations. Simulations studies were conducted with $\delta = 1$ and the mean of Y was computed assuming MCAR, MI assuming MAR, and their new proposed re-weighting MAR for MNAR imputations. For each increasing number of imputations, the re-weighting approach outperformed the other methods. However, a drawback of their approach is that they do not take into account uncertainty regarding the missing data mechanism and they did not investigate the impact of misspecifying the propensity of missingness δ. Also the the standard errors associated with their method may generally be underestimated since the degrees of freedom for the t-distribution of the MAR imputation should be decreased for the NMAR estimator because re-weighting decreases the effective sample size. A consequence of underestimated standard errors is confidence intervals which do not have desired coverage rates.

Siddique et al. (2012) developed an imputation procedure that actually incorporates missing data mechanism uncertainty by specifying a range of ignorability assumptions and combining these assumptions into one inference. The I models are drawn from the distribution of nonignorable (and ignorable) models by multiplying the ignorable imputed values by a factor of k defined as follows: (nonignorable imputed Y_i) $=$ $[(k-1) \times |$ignorable imputed $Y_i|] +$ ignorable imputed Y_i. The ignorable imputed Y_i are generated by regression imputation (Rubin, 1987) and the multiplier k is formed by multiple draws from a distribution which depends on the ignorability assumptions. For example, specifying k as Normal (1.5, 0.5) suggest that the imputer believes that missing values tend to be larger than the observed values thus relying on a non-ignorability assumption. Then J multiple imputations are made for each of the i models which results in $I \times J$ completed data sets. More specifically, define ψ as the imputation model which is drawn from the predictive distribution

$$\psi^i \sim p(\psi), \text{ where } i = 1, ..., I.$$

Then j independent imputations conditional on ψ^i are drawn such that an imputation for the missing components of Y, known as Y_{mis} is

$$Y_{mis}^{(i,j)} \sim p(Y_{mis}|Y_{obs}, \psi^i) \text{ where } j = 1, ..., J$$

where Y_{obs} is the observed units of Y. However, the $I \times J$ nested multiple imputations are not independent draws from the same posterior predictive distribution of Y_{mis}, therefore Rubin (1987) combining rules for multiply-imputed data sets were not applicable. They adopt combining rules for the nested multiple imputations which generates an overall average of the $I \times J$ point estimates, overall average of the associated variance estimates, the within-model variance, and the between-model variance. The multiple-model approach of Siddique et al. (2012) is an improvement over methods that make no assumptions regarding missing data mechanism uncertainty. Although in real world settings it may be impossible to know precisely the degree of nonignorablity, hence the distribution of k. Our new proposed multiple imputation approach alleviates the need to specify unverifiable adjustment parameters such as k. We propose a sensitivity analysis which allows users to incorporate prior knowledge on the missing data model and response model through the use of weights. Our approach allows the use of auxiliary variables that are not part of the analysis procedure to be incorporated into the response model to decrease bias and increase efficiency.

Kim and Kim (2012) proposed an imputation method for general purpose estimation based on the parametric fractional imputation for nonignorably missing data. The methodology involves using a fraction of all observed values defined in the corresponding imputation cell to impute the unobserved values. Similar to the approach of Carpenter et al. (2007) weights are derived using importance sampling and the Expectation Maximization (EM) algorithm, but instead of being applied to adjust the MAR estimates, it is applied as a fraction of the observed values. Under nonignorable missingness, the parameters associated with the outcome model β and the nonresponse model α must be estimated simultaneously. To find the estimator that maximizes the observed likelihood $L_{obs}(\alpha, \beta)$, which equates to solving the mean score function for α and β, we set $\bar{S}(\alpha, \beta|\hat{\alpha}_{(t)}, \hat{\beta}_{(t)}) = 0$, where t is the t-th iteration in the EM algorithm and the fractional weights are defined as

$$w_{ij(t)} \propto \frac{f_1(y_i^{*(j)}|\boldsymbol{x}_i, \hat{\beta}_{(t)})\{1 - \pi(\boldsymbol{x}_i, y_i^{*(j)}; \hat{\alpha}_{(t)})\}}{\hat{f}(y_i^{*(j)}|\boldsymbol{x}_i, R_i = 1)}$$

with $\sum_{j=1}^{M} w^*_{ij(t)} = 1$ and $\hat{f}(y_i^{*(j)}|\boldsymbol{x}_i, R_i = 1)$ which is the estimated conditional distribution used to generate the imputed values for $Y_{i,mis}$. The authors suggest alternatives for avoiding really large fractional weights by using a different form of imputing values. Also an easier approach that does not rely on the EM can be implemented if follow-up data are available. Furthermore, if the number of imputations M is large, then a calibration weighting technique can be used to calculate the fractional weights. Variance estimation can be computed for the parametric fractional imputation using a replication method such as jackknife or bootstrap. Simulation results were shown to demonstrate that the resulting estimator is very close to the maximum likelihood estimator. Kim and Kim (2012) believe their approach has an advantage over multiple imputation. Fractional imputation uses a fraction of all the observed values in the imputation cell and therefore there is no need to add imputation variability which can lead to greater efficiency than the multiple imputation estimator. However, it is difficult to compare the efficiency of MI and parametric fractional imputation because the variances are defined differently.

Another parametric, sensitivity-free method for imputing nonignorable missing data is the Random Indicator (RI) method introduced by Jolani (2012). The basic principal involves generating a pseudo response indicator by drawing random values from the model for the missingness and iteratively imputing the incomplete variable. An initial imputation for the incomplete variable $Y^{(0)} = (Y_{obs}, Y_{mis}^{(0)})$ is created assuming the data are MAR. Suppose data are also available on a fully observed covariate $\boldsymbol{X}$, then a pseudo response indicator $\dot{R}$ is drawn from a Bernoulli process $\dot{R} \sim \text{Bernoulli}(1, \pi)$ where $\pi \sim P(R = 1|X, Y^{(t)})$ and t is an iteration. Based on the assumption of normality, the expectation for the imputation model were presented as the following set of equations,

$$\begin{aligned} E(Y|R = 1) &= \boldsymbol{X}\beta - \delta_R(1 - \dot{R}) \quad (1.2) \\ E(Y|R = 0) &= \boldsymbol{X}\beta - \delta_{NR}(2 - \dot{R}) \quad (1.3) \end{aligned}$$

where δ_R is the adjustment parameter estimated from the observed part of the data (eq 1.2) and the missing data are imputed using δ_{NR} (eq 1.3). The assumption is $\delta_R = \delta_{NR}$ and is shown to be equivalent by proofs that rely on the assumption of normality of Y_{obs} which can be tested. The missing data are predicted for the case that $R = 0$ and $\dot{R} = 1$ using $\boldsymbol{X}\dot{\beta} - \hat{\delta}_{NR}$ and for the case that $R = 0$ and $\dot{R} = 0$ using $\boldsymbol{X}\dot{\beta} - 2\hat{\delta}_{NR}$ where $\dot{\beta}$ is drawn from its posterior distribution for a given prior for β. The iterative algorithm runs between the following two steps of predicting Y_{mis} and $\dot{R}$ until convergence. In the last stage of the imputations, an adjustment must be applied to add the appropriate amount of noise to the predicted imputation values to account for the missing values that are more likely to be smaller or larger. Although a novel approach, the RI method was restricted to the assumption that Y_{obs} comes from a normal distribution and the variance of the responders was equivalent to the variance of the nonresponders. The method also produced larger standard errors than the complete-case method which leads to less efficiency.

Sullivan and Andridge (2015) also proposed a parametric approach to mutliply impute nonignorable missing data through the use of the hot deck. They propose a proxy pattern-mixture hot deck which creates a proxy for all nonrespondents and bootstrapped respondents by regressing Y on fully observed covariates X. Predicted values for Y are based on a pattern-mixture model and nonrespondent values are varied by a sensitivity parameter λ that determines the missingness mechanism. The values λ are varied as 0 to assume a MAR mechanism, 1 for a weak MNAR mechanism, and $\lambda = \infty$ assuming an extreme case of MNAR. The limitations of this method are that the parameter λ is hard to interpret beyond MAR and MNAR and its not quite clear that the method can be extended to general missing data patterns because only the observed cases are bootstrapped in the beginning of the procedure. In a subsequent paper, the method is extended to account for binary data.

1.3.3.2 Nonparametric Imputation Methods for MNAR data

In addition to the parametric multiple imputation procedure, some nonparametric methods have been proposed through the use of the approximate Bayesian bootstrap (ABB) and hot deck. Multiple imputation using the ABB with follow-up data for a random sample of the nonrespondents was proposed in Glynn et al. (1993). The ABB was performed by first randomly drawing from the sample of followed-up nonrespondents followed by a random draw from the same sample for the nonrespondents with no follow-up data. The nonrespondents values were then imputed M times to create M completed data sets. However, there are serious limitations of this approach considering it relies on complete data for a random sample of the nonrespondents which may be very difficult or costly to obtain in practice.

Siddique and Belin (2008) proposed a hot deck method to multiply impute nonignorably missing data through the use of the hot deck. The basic principal of hot deck imputation is to use a respondent's (donor) observation to impute the missing values in the nonrespondents (recipients). An ignorable ABB draws the observed cases randomly with replacement from Y_{obs} to create Y^*_{obs}. As contrasted with the ignorable

ABB, in the nonignorable ABB the bootstrap observed cases Y^*_{obs} of Y_{obs} are drawn with probability of selection proportional to Y^c such that $y_i \in Y_{obs}$ is $\frac{y_i^c}{\sum_{j=1}^{n_{obs}} y_j^c}$. The $c = \{-1, 0, 1, 2, 3\}$ and is chosen depending on the assumption of Y_{mis}. For example, if one assumes the nonrespondents have larger values than the respondents then $c = 3$; likewise, $c = -1$ if one assumes the nonrespondents have smaller values. The $c = 0$ implies an ignorable hot deck where imputations for a nonrespondent are randomly drawn with replacement from the respondent values. Once bootstraps samples are generated from the observed cases to ensure a 'proper' imputation (Rubin, 1987), then a regression of the bootstrap samples on the fully observed variables is performed. Predicted values are formed for all the nonrespondents $\hat{Y}_{NR}$ and for the respondents $\hat{Y}_R$ in the bootstrapped sample. A distance-based donor selection procedure was introduced with distances defined as $D^k_{NR,R} = (|\hat{y}_{NR} - \hat{y}_R| + \delta)^k$, where δ is a nonzero offset that is the minimum distance between $\hat{y}_{NR}$ and $\hat{y}_R$. The closeness parameter k adjusts the probability of selection assigned to the closest donors and as $k \to \infty$ then this procedure reduces to a nearest-neighbor hot deck. If the exponent $k = 0$, then it is equivalent to a simple random hot deck. The authors suggested that $k = 3$ had the most favorable results with respect to favoring nearby donors. In this method, covariate information is used to obtain predicted values but is not incorporated in correction for nonignorability. As a consequence, covariate information is lost. We propose a more robust, nonparametric MI method that harnesses the power of bootstrapping and the hot deck but is distinctive from these methods. Our method uses a model for the data and also incorporates a model for the probability of missingness by using available covariate information in the data model and the probability of missingness model.

1.4 Diagnostic Methods for Missing Data Models

Multiple imputation is a popular method for handling missing data. However, it is not common in practice to perform diagnostic checks to determine whether the imputations formed from an imputation model are plausible. Meng (1994a) suggested that the imputation model be congenial or general enough to preserve any associations among variables that may be the target in the subsequent, completed data analyses. Furthermore, a general imputation model that is close to the true model allows for accommodation of a wide range of statistical models that can be used on the completed data sets. In order to construct a reasonable, general imputation model, a major issue is to not exclude any important predictors or relationships (i.e. nonlinear relationships) among the data. Excluding important features may lead to imputation models that are not as general than the subsequent analysis and can potentially bias the results. Diagnostic and model checking of imputation models is a natural way to determine whether these assumptions hold.

Although diagnostic methods are scarce, diagnostic testing for missing data models have a long history. One of the first studies for diagnostic for missing values was introduced by Poirier and Ruud (1983). They introduced a more general case of the Heckman (1977) selection model to handle missing data in a maximum likelihood based approach. Violation of either homoscedasticity and lognormality could potentially result in inconsistency of the estimators in those models. To identify departures from model assumptions, they proposed the use of Lagrange multipliers test. However, they did not study diagnostic for imputations because it was before (Rubin, 1987) published his pioneering work on MI to handle nonresponse in surveys.

Previous work on imputation diagnostic is limited to the assumption that the data are missing at random. For instance, Raghunathan and Bondarenko (2007) proposed the use of propensity scores as a diagnostic tool to check the validity of imputed vales in MI. They checked the equality of the distributions of the observed and missing values conditional on the response propensity score. Wang (2010) extended this diagnostic approach to include a regression of both the observed and imputed data as a function of the predicted propensity score and the missingness indicator. With the extension, Wang (2010) was able to check whether the imputation model used to generate imputations would preserve the associations among variables in the dataset by determining if the missingness was completely explained by the response propensity score.

Several authors have used graphical tools and numerical test such as Kolmogorov-Smirnov test to assess plausibility of imputations. For example, Abayomi et al. (2008) examined the empirical density plots, bivariate scatter plots, and residual plots to identify dramatic differences from the observed and imputed data. Bondarenko and Raghunathan (2016) made graphical comparisons of the observed and imputed values conditional on the response propensity to assess the suitability of imputations from imputation models. Abayomi et al. (2008) and Nguyen et al. (2013) used the KS test to diagnose problems with imputation models by comparing the empirical distribution of the observed and imputed data. Abayomi et al. (2008) flagged imputed variables with statistically significant differences and further examined variables using graphical techniques. Nguyen et al. (2013) suggested that the imputed variables required more rigorous evaluation after the KS test is performed. Nguyen examined the behavior of the KS p-value under various scenarios in simulations, including varying the amount of missing data, misspecified imputation models, and skewed and heavy-tailed distributions.

He and Zaslavsky (2012) and Nguyen et al. (2015) used a diagnostic method based on posterior predictive checking [PPC] (Gelman et al., 1996), namely the posterior predictive p-value (Meng, 1994b), to determine the adequacy of imputation models by applying subsequent analyses of interest to both the completed data and their posterior replicates simulated under the imputation model. Large differences between the estimates using the completed data and the simulated replicates may suggest model inadequacy. He and Zaslavsky (2012) and Nguyen et al. (2015) checked imputation models assuming the missing data are MAR. However, principled diagnostic approaches that can handle the less restrictive assumption of MNAR have not been investigated in the literature. Motivated by such facts, our primary goal is to evaluate whether the diagnostic methods of He and Zaslavsky (2012) and Nguyen et al. (2015) are applicable in the case of MNAR imputation models.

1.5 Motivating examples

Throughout this dissertation, we use two data sets with variables with missing values. The first data set is from a prostate cancer study. These data are available in the Gene Expression Omnibus (GEO) database under accession number GDS3289. It contains 104 samples, including 34 benign epithelium samples and 70 non-benign samples. Missing values are present for some genomic biomarkers. There are 1894 biomarkers that do not have missing values. However, 18,111 variables have missing values.

The second data set is from the Georgia Coverdell Acute Stroke Registry. A detailed description of the registry data set was reported elsewhere (Camp et al., 2015). The primary goals of the registry are to gain a better understanding of factors associated with stroke and improve the quality of acute stroke care of patients. There are over 86322 subjects and over 203 variables of which 135 have missing values with a general missing data pattern.

1.6 Outline

In this dissertation, we present robust statistical methods for various assumptions on missing data. In Chapter 2, we develop multiple imputation methods to handle missing data in the presence of high-dimensional where data are assumed to be missing at random. In Chapter 3, we present nonparametric imputation methods to handle nonignorable missing data by using screening, sparse principal component analysis, and sufficient dimension reduction techniques. Finally, in Chapter 4 we examine whether posterior predictive checking is applicable for imputation diagnostics under nonignorable missingness.

Chapter 2

Multiple Imputation using Dimension Reduction Techniques for High-Dimensional Data

Abstract

Missing data present challenges in data analysis. Naive analyses such as complete-case and available-case analysis may introduce bias and loss of efficiency, and produce unreliable results. Multiple imputation (MI) is one of the most widely used methods for handling missing data which can be partly attributed to its ease of use. However, existing MI methods implemented in most statistical software are not applicable to or do not perform well in high-dimensional settings where the number of predictors is large relative to the sample size. To remedy this issue, we develop an MI approach that uses dimension reduction techniques. Specifically, in constructing imputation models in the presence of high-dimensional data our approach uses sure independent screening followed by either sparse principal component analysis (sPCA) or sufficient dimension reduction (SDR) techniques. Our simulation studies, conducted for high-dimensional data, demonstrate that using SIS followed by sPCA to perform MI achieves better performance than the other imputation methods including several existing imputation approaches. We apply our approach to analysis of gene expression data from a prostate cancer study.

2.1 Introduction

Appropriate handling of missing data requires an understanding of its source and structure. It is well known that naive analyses such as complete-case and available-case analysis may introduce bias and loss of efficiency, and produce unreliable results. Multiple imputation (MI) (Rubin, 1976, 1987) is one of the most widely used methods which can be partly attributed to its ease of use. The basic idea underlying MI is to replace missing values M times by "plausible values" drawn from their posterior predictive distributions given the observed data. Multiply imputed data sets are generated to account for sampling variability and uncertainty of imputing missing values. Then each data set completed by imputation is analyzed using the standard complete-data methods and the estimates obtain from these analyses are combined using Rubin's rule (Rubin, 1987) to create one statistical inference summary. A key advantage of MI is that the imputation model can be operationally distinct from the subsequent analyses (target analysis that would be performed in the absence of missing data). The use of MI has been investigated in various settings and detailed reviews are provided elsewhere. (Harel and Zhou, 2007; Carpenter and Kenward, 2012)

The Problem

The validity of MI is predicated on several assumptions. First, the missing at random (MAR) (Little and Rubin, 2014) mechanism is often assumed and implies the missingness is not associated with the missing values conditional on observed data.(Rubin, 1976) Our current work assumes that the incomplete data are MAR. Second, Meng (Meng, 1994a) suggested that the imputation model be congenial or general enough to preserve any associations among variables that may be the target in the imputed data analyses. Furthermore, a general imputation model that is close to the true model allows for accommodation of a wide range of statistical models that can be used on the imputed data sets. In order to construct a reasonable, general imputation model, a major issue is to not exclude any important predictors, since excluding important variables may lead to imputation models that are not as general than the subsequent analysis and will potentially bias the results. However, in practice it is not feasible to specify all possible relevant predictors and their interactions in an imputation model. A more challenging problem arises in the presence of high-dimensional data where the number of variables is larger than or approximately equal to the sample size. Largely due to the advancement of technology, the amount of data collected is rapidly increasing which give rise to high-dimensional data. Examples of high-dimensional data include genomics, proteomics and functional magnetic resonance imaging data. These data often contain missing values, yet there has been limited work in developing approaches for handling missing data in the presence of high-dimensional data. Standard MI approaches implemented in most statistical software perform poorly or fail in the presence of high-dimensional data.(Zhao and Long, 2013a)

Existing approaches for MI in the Presence of High-Dimensional Data

Model trimming is essential to construct imputation models in the presence of high-dimensional data. Stekhoven and Bühlmann (2012) used a classification technique, namely random forest (RF), to impute missing values in high-dimensional data. The variable with missing values is treated as the response variable and other (auxiliary) variables are used for bootstrap aggregation of multiple regression trees to potentially reduce overfitting. The predictions are combined from trees to improve accuracy of prediction of the missing values. However, the selection of tuning parameters such as the number of trees and number of nodes needs further investigation. Liao et al. (2014) proposed another imputation approach for high-dimensional data which is a variation of a K-nearest-neighbor imputation. For a missing value, the method seeks its K nearest variables (KNN_V) or subjects (KNN_S) and imputes by a weighted average of observed values of the similar neighbors. Although the method was shown to perform well in their simulations where the performance was evaluated based on comparisons between true and imputed values, it does not properly propagate the uncertainty in estimating the parameters in the imputation model and hence it is not proper in the sense of Rubin (Rubin, 1987). Improper imputation can lead to biased inference in the subsequent analyses.

Zhao and Long (2013a) proposed an MI approach for high-dimensional data based on regularized regression that does account for the uncertainty in imputation. Specifically, they investigated the use of MI through direct and indirect use of regularized regression. In the former, regularized regression is used for both variable selection and parameter estimation for imputation models; in the latter regularized regression is only used for model trimming. Direct use of regularized regression in MI was shown to achieve superior performance in the settings considered in their work. They also proposed an MI method using the Bayesian lasso (Blasso)(Park and Casella, 2008) to estimate and select important variables in imputation models. However, these methods also have some limitations and particularly they may not yield good performance when the true imputation model is large. To tackle this challenge, we consider an alternative approach to constructing imputation models by incorporating dimension reduction techniques.

2.1.1 Dimension Reduction Techniques for High-Dimensional Data

Screening is an effective strategy to deal with high dimensionality. In particular, sure independent screening (SIS) (Fan and Lv, 2008) is a method which is based on correlation learning which filters out the features that have weak correlation with the response. Another dimension reduction technique is sparse principal component analysis (sPCA).(Zou et al., 2006) The commonly used principal component analysis (PCA) seeks linear combinations of p variables such that the derived components capture maximum variance. Yet, a drawback of PCA is that the loadings of all p variables are typically nonzero, which is often hard to interpret. Zou et al. (2006) modified PCA by using the lasso penalty (sPCA_ST) to shrink some loadings to zero, allowing for identification of important features. More recently, other authors proposed adjustments to sparse principal component analysis. Witten et al. (2009) used penalized matrix decomposition (sPCA_PMD), a regularized version of the singular value decomposition, to create sparse loadings. Lee et al. (2012) proposed two approaches to modify sPCA by using the lasso (Tibshirani, 1996) (sPCA_L) and adaptive lasso (Zou, 2006) (sPCA_AL) penalty terms.

Alternatively, we can use sufficient dimension reduction regression (Weisberg, 2002) (SDR) to find relevant predictors in imputation models. SDR seeks to find d linearly independent linear combinations such that all the information about the regression is contained in the d linear combinations and d is typically considerably less than the number of variables (namely, p). There are several variations of SDR including sliced inverse regression (SIR) (Li, 1991), sliced average variance estimates (SAVE) (Dennis Cook, 2000), and principal Hessian directions (PHD) (Li, 1992).

In this paper, we propose a new MI approach that imputes the missing values by first screening for relevant predictors of a variable with missing values. Using the screened variables, we further reduce dimensions by applying SDR or sPCA and use the resulting linear combinations to construct imputation models. The remainder of this paper is organized as follows. In the next section, we describe the proposed approach based on SPCA and SDR. In the following section, we perform simulations to evaluate the performance of the proposed approach in comparison with several existing approaches including Blasso, DAlasso, KNN_S, KNN_V, and RF in the presence of high-dimensional data. In the fourth section we illustrate the new proposed approach using genomics data from a prostate cancer study. We conclude with a discussion in the last section.

2.2 Methodology

To fix ideas, let $\boldsymbol{Y}$ denote a set of p variables observed for a random sample of n observations. Denote by $\boldsymbol{Y}_{obs}$ the observed components of $\boldsymbol{Y}$ and by $\boldsymbol{Y}_{mis}$ the missing components of $\boldsymbol{Y}$. Suppose that $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ follows a model $\pi(\boldsymbol{Y}|\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is a set of parameters and the missing data mechanism is missing data at random (MAR). Under ignorability, the standard imputation framework can be represented by (2.1)

$$\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \int \pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{Y}_{obs})d\boldsymbol{Y}. \tag{2.1}$$

Specifically, one can first generate a random draw from the posterior distribution of $\boldsymbol{\beta}$

$$\boldsymbol{\beta}^{(m)} \sim \pi(\boldsymbol{\beta}|\boldsymbol{Y}_{obs}),$$

and then generate a random draw of the missing values from their posterior predictive distributions

$$\boldsymbol{Y}_{mis}^{(m)} \sim \pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\beta}^{(m)}),$$

where $m = 1, ..., M$ and M is the number of imputed data sets. (Schafer, 1999)

For ease of exposition, we describe our proposed approach in a setting where only one variable $\boldsymbol{y}_1$ has missing values with the remaining variables $\{\boldsymbol{y}_2, ..., \boldsymbol{y}_p\}$ fully observed, and all $\boldsymbol{y}$ are continuous variables. Let n_1 denote the number of complete cases with all variables observed and n_2 the number of incomplete cases with y_1 missing ($n = n_1 + n_2$). Define $\boldsymbol{y}_{obs,1} = (y_{1,1}, y_{2,1}, ..., y_{n_1,1})^T$ as the first n_1 observed components of $\boldsymbol{y}_1$ and its complement as $\boldsymbol{Y}_{obs,-1} = (\boldsymbol{y}_{1,-1}, \boldsymbol{y}_{2,-1}, ..., \boldsymbol{y}_{n_1,-1})^T$ with $\boldsymbol{y}_{i,-1} = (y_{i,2}, y_{i,3}, ..., y_{i,p})$, which together form the set of complete cases. Define $\boldsymbol{y}_{mis,1} = (y_{n_1+1,1}, y_{n_1+2,1}, ..., y_{n,1})$ as the $n - n_1$ missing components of $\boldsymbol{y}_1$ and its complement as $\boldsymbol{Y}_{mis,-1} = (\boldsymbol{y}_{n_1+1,-1}, \boldsymbol{y}_{n_1+2,-1}, ..., \boldsymbol{y}_{n,-1})$, which together form the set of incomplete cases. Of note, $\boldsymbol{Y}_{mis,-1}$ is observed. It follows that the observed data are $(\boldsymbol{y}_{obs,1}, \boldsymbol{Y}_{obs,-1}, \boldsymbol{Y}_{mis,-1})$ and the missing data are $\boldsymbol{y}_{mis,1}$. The imputation model (2.1) reduces to

$$\pi(\boldsymbol{y}_{mis,1}|\boldsymbol{y}_{obs,1}, \boldsymbol{Y}_{obs,-1}, \boldsymbol{Y}_{mis,-1}) = \int \pi(\boldsymbol{y}_{mis,1}|\boldsymbol{Y}_{mis,-1}, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{y}_{obs,1}, \boldsymbol{Y}_{obs,-1})d\boldsymbol{\beta}. \quad (2.2)$$

To complete the imputation model (2.1), we can posit a regression model with $\boldsymbol{y}_1$ as the outcome

$$\boldsymbol{y}_1 = \delta_0 + \boldsymbol{Y}_{obs,-1}\boldsymbol{\delta} + \epsilon \quad (2.3)$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2 I_{n_1})$ and $\boldsymbol{\beta} = (\delta_0, \boldsymbol{\delta}, \sigma^2)^T$. Model (2.3) can be fitted using the set of complete cases. However, when $p \gg n$, standard regression techniques such as ordinary least squares fail and it is imperative to perform variable selection or dimension reduction when fitting model (2.3). As demonstrated in our simulations, when the true model for (2.3) is large, i.e., the number of important predictors in (2.3) is large relative to n, imputation methods based on regularized regression may yield unsatisfactory performance.

We propose to use dimension reduction techniques when constructing imputation models, specifically, applying SIS followed by either sPCA or SDR before fitting model (2.3). The proposed imputation approach is detailed as follows:

1. In the first step, SIS is performed using the complete cases to find a subset of v variables that are predictive of the incomplete variable y_1. Let $\{t_1, ..., t_v\}$ index the subset of v variables selected from $y_2, \ldots, y_p$ using SIS, where $v < p - 1$.

2. In the second step, we achieve further dimension reduction via either sPCA or SDR, as v can be still large relative to n.

   (a) Applying sPCA to the set of v variables selected in the first step and using all n observations, we obtain

$$\begin{aligned}
\boldsymbol{z}_1 &= \alpha_{1,1}\boldsymbol{y}_{t_1} + \alpha_{1,2}\boldsymbol{y}_{t_2} + ... + \alpha_{1,v}\boldsymbol{y}_{t_v} \\
\boldsymbol{z}_2 &= \alpha_{2,1}\boldsymbol{y}_{t_1} + \alpha_{2,2}\boldsymbol{y}_{t_2} + ... + \alpha_{2,v}\boldsymbol{y}_{t_v} \\
&\vdots \\
\boldsymbol{z}_v &= \alpha_{v,1}\boldsymbol{y}_{t_1} + \alpha_{v,2}\boldsymbol{y}_{t_2} + ... + \alpha_{v,v}\boldsymbol{y}_{t_v}
\end{aligned}$$

   where the linear combinations $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_v$ are the principal components, and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_v$ are the loading vectors. We select $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, ... , $\boldsymbol{z}_d$ in sPCA by either choosing the first principal component or the first d principal components that explain at least 60 or 80% of the total variance. It is important to note that d is typically substantially less than the original number of variables p and the sample size n.

   (b) Alternatively, applying SDR to the set of v variables selected in the first step, we obtain

$$\begin{aligned}
\boldsymbol{z}_1 &= \gamma_{1,1}\boldsymbol{y}_{t_1} + \gamma_{1,2}\boldsymbol{y}_{t_2} + ... + \gamma_{1,v}\boldsymbol{y}_{t_v} \\
\boldsymbol{z}_2 &= \gamma_{2,1}\boldsymbol{y}_{t_1} + \gamma_{2,2}\boldsymbol{y}_{t_2} + ... + \gamma_{2,v}\boldsymbol{y}_{t_v} \\
&\vdots \\
\boldsymbol{z}_d &= \gamma_{d,1}\boldsymbol{y}_{t_1} + \gamma_{d,2}\boldsymbol{y}_{t_2} + ... + \gamma_{d,v}\boldsymbol{y}_{t_v}
\end{aligned}$$

where $\boldsymbol{y}_1$ is used as the response variable, γ's are the estimated coefficients in SDR using the set of complete cases, and d $(d < v)$ is chosen by using an asymptotic test for PHD and permutation tests for SAVE and SIR (Weisberg, 2002). The most notable difference between sPCA and SDR is that SDR uses the variable with missing values, namely $\boldsymbol{y}_1$, as the outcome to guide dimension reduction.

After obtaining $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, ... , $\boldsymbol{z}_d$ via either sPCA or SDR, define $\boldsymbol{Z}_{obs} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_{n_1})^T$ for the complete cases and $\boldsymbol{Z}_{mis} = (\boldsymbol{z}_{n_1+1}, \boldsymbol{z}_{n_1+2}, ..., \boldsymbol{z}_n)$ for the incomplete cases. Of note, both $\boldsymbol{Z}_{obs}$ and $\boldsymbol{Z}_{mis}$ can be calculated, since they involve only a subset of $y_2, \ldots, y_p$.

3. In the third step, we replace $\boldsymbol{Y}_{obs,-1}$ and $\boldsymbol{Y}_{mis,-1}$ with $\boldsymbol{Z}_{obs}$ and $\boldsymbol{Z}_{mis}$ respectively in (2.2) and (2.3) and conduct imputation accordingly. More specifically, using model (2.3) with $\boldsymbol{y}_{obs,1}$ as the outcome variable and $\boldsymbol{Z}_{obs}$ as predictors, we randomly draw $\hat{\boldsymbol{\beta}}^{(m)}$ from its posterior distribution and then impute $\boldsymbol{y}_{mis,1}$ using $\boldsymbol{Z}_{mis}$ by drawing randomly from the conditional posterior predictive distribution $\pi(\boldsymbol{y}_{mis,1}|\boldsymbol{Z}_{mis}, \hat{\beta}^{(m)})$, where $m = 1, \ldots, M$ for M imputations.

Once the missing data are multiply imputed, subsequent analyses such as multiple regression or logistic regression are performed for each of the M imputed datasets. Analysis results are then combined for statistical inference using Rubin's combining rule (Rubin, 1987).

2.3 Simulation studies

We conduct simulation studies to evaluate the performance of our proposed approach. In the second step of our proposed approach, we use either sPCA or SDR. For sPCA, we consider four variations, namely, sPCA_ST (Zou, 2006), sPCA_PMD (Witten et al., 2009), sPCA_L (Lee et al., 2012), and sPCA_AL (Lee et al., 2012). The sPCA_ST and sPCA_PMD methods are both implemented in R package **PMA**(Witten et al., 2013). The sPCA_L and sPCA_AL methods are both implemented in R code provided on the authors website.(Lee et al., 2012) For SDR, we consider three variations, namely, sliced average variance estimates (Dennis Cook, 2000) (SDR_SAVE), sliced inverse regression (Li, 1991) (SDR_SIR), and principal Hessian directions (Li, 1992) (SDR_PHD). The SDR methods are implemented in the R package **dr**(Weisberg, 2002). We compare our approach to several existing imputation methods proposed by Zhao and Long (2013a), Liao et al. (2014), and Stekhoven and Bühlmann (2012). Zhao and Long (2013a) used Bayesian lasso regression (Blasso) and adaptive lasso with direct use of regularized regression (DAlasso) to conduct MI in high-dimensional data. Liao et al. (2014) used variations of k-nearest neighbors, namely, KNN_S and KNN_V for imputation of missing values. Stekhoven and Bühlmann (2012) proposed random forest (RF) for MI. We also include the standard parametric MI procedure implemented in the R package **mice**(van Buuren and Groothuis-Oudshoorn, 2011) with the default method of Bayesian linear regression.

In our simulations we focus on estimating the regression coefficients $\hat{\theta}$ from linear regression in the presence of missing data. We use $\hat{\theta}$ obtained from the imputed data sets to compare the performance across different imputation methods. To have a point of reference for bias and efficiency for estimating θ, we apply a gold standard (GS) method that estimates θ using the underlying complete data before missing data are generated. We also perform a complete-case analysis (CC), in which only the set of complete cases are used in data analysis.

2.3.1 Simulation setup

We vary several factors including the total number of variables (p), the number of variables in the true imputation model (c), and the correlation among the data (ρ). Simulations are carried out with 500 Monte Carlo (MC) datasets and the sample size is fixed at $n = 100$ in each MC dataset. Each simulated data set includes the fully observed outcome variable ($\boldsymbol{w}$) and the set of predictors and auxiliary variables $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_p)$. The variable $\boldsymbol{y_1}$ contains missing values $\boldsymbol{y}_{mis,1} = (y_{n_1+1,1}, y_{n_1+2,1}, ..., y_{n,1})$. The details of the simulation set-up are as follows:

I. $\boldsymbol{Y_{-1}} = (\boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_p)$ is generated from a multivariate normal distribution with mean $(0, 0, ..., 0)_{p-1}$ and a first-order autoregressive covariance matrix with autocorrelation, denoted by ρ, varying as 0.1, 0.5, or 0.9. We also consider a block diagonal covariance matrix having main diagonal blocks square matrices with the off-diagonal blocks as zero matrices. The main diagonal block matrices are composed of compound symmetric matrices with $\sigma^2 = 2$ on the diagonals and $\sigma^2\rho$ on the off-diagonals, where ρ is again varied as 0.1, 0.5, and 0.9. We consider settings with $p = 200$ and $p = 1000$.

II. For each combination of p and ρ, $\boldsymbol{y}_1$ is generated from a normal distribution such that $y_1 \sim \mathrm{N}(\boldsymbol{Y}_T\boldsymbol{\eta}, 1)$, where T represents the set of the variables in the true imputation model with a cardinality of c. Two cases are considered with the corresponding design matrices $\boldsymbol{Y}_T = (\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_{50}, \boldsymbol{y}_{51})$, $(\boldsymbol{y}_2, ..., \boldsymbol{y}_{51}, \boldsymbol{y}_{100}, ..., \boldsymbol{y}_{149})$ for $c = 4$ and $c = 100$, respectively. We set $\boldsymbol{\eta} = \boldsymbol{1}' \times 1, \boldsymbol{1}' \times 0.05$ to include the intercept for $c = 4$ and $c = 100$, respectively. The values of $\boldsymbol{\eta}$ are chosen to fix the signal-to-noise ratio when generating y_1. Of note, $c = 100$ corresponds to the case where the size of the true imputation model is large relative to the sample size.

III. Given $\boldsymbol{Y}, \boldsymbol{w}$ is generated from a normal distribution $w \sim \mathrm{N}(\theta_0 + \theta_1 y_1 + \theta_2 y_2 + \theta_3 y_{10}, \sigma_2 = 3)$, where all $\theta = 1$.

IV. The response indicator R for $\boldsymbol{y}_1$, which is 1 if $\boldsymbol{y}_1$ is observed and 0 otherwise, is generated from a logistic model, $\mathrm{logit}[\Pr(R = 1|\boldsymbol{Y}_{-1}, \boldsymbol{w})] = -1 - 0.1\boldsymbol{y}_2 + 2\boldsymbol{y}_3 - 10\boldsymbol{w}$ which results in an average of 31% of $\boldsymbol{y}_1$ missing.

We conduct imputation of $y_{mis,1}$ using each imputation approach considered. The subsequent analysis for the imputed data is performed using a linear regression of $\boldsymbol{w}$ on the imputed $\boldsymbol{y}_1$ and fully observed $\boldsymbol{y}_2$ and $\boldsymbol{y}_{10}$. For our proposed method and the existing MI based methods, we use the R package **mice** to multiply impute the missing data with its default method of Bayesian linear regression with a ridge parameter value of 10^-5. We also investigate the use of ridge parameter values 10^{-1} and 10^{-3}. Thirty imputed data sets are generated for each MI method to use in subsequent analyses. Rubin's rule (Rubin, 1987) is used to pool the estimates to obtain $\hat{\boldsymbol{\theta}}$ and their standard errors.

2.3.2 Results

The simulation results are summarized for $\hat{\theta}_1$ which is the parameter estimate that is associated with the incomplete variable $\boldsymbol{y}_1$. Tables 2.1, 2.2, 2.3, and 2.4 present the mean bias of $\hat{\theta}_1$ (Bias), mean standard error of $\hat{\theta}_1$ (SE), Monte Carlo standard deviation of $\hat{\theta}_1$ (SD), mean square error of $\hat{\theta}_1$ (MSE), and coverage rate of the 95% confidence interval of $\hat{\theta}_1$ (CR). In addition to comparing different methods, we evaluate the effect of the dimension of the data (p), the number of variables in the true imputation model (c), and the correlation among the data $\boldsymbol{Y_{-1}}$ (ρ). Within each table, p and ρ are varied while c is fixed. In both Tables 2.1 and 2.2, the covariance structure is an autoregressive matrix. More specifically, $c = 4$ in Table 2.1 and $c = 100$ in Table 2.2. In Table 2.3 and Table 2.4, the covariance structure is a block diagonal matrix with compound symmetric blocks. We let $c = 4$ in Table 2.3 and let $c = 100$ in Table 2.4. In our proposed approach, we use the R package **SIS** (Fan et al., 2014) with the default vanilla method to find a subset of v variables that are predictive of the incomplete variable y_1. In each setting, SIS selects between 10 and 17 variables. SPCA is conducted using four variations and the number of principal components (PCs) are selected by either choosing the first principal component or the first d principal components that explain at least 60 or 80% of the total variance. We observe superior performance using one principal component.

In comparing the existing methods to our new proposed methods, we observe that our new proposed method outperforms the existing methods. When the size of the true imputation model is small ($c = 4$) relative to n, the Blasso imputation method of Zhao and Long (2013a) yields modest bias. In contrast, the CC, MI, DAlasso, KNN and RF methods, in general, yield substantial bias and inadequate CRs. More importantly, our methods outperform all the existing methods, including Blasso, in terms of bias, MSE, and coverage rates. Furthermore, as c increases to 100, there is considerably more pronounced deterioration in the performance of Blasso compared to our proposed approach. Our proposed method has minimal bias, coverage rates near the nominal level, and overall superior performance as compared to the existing methods, irrespective of whether the true imputation model is small or large relative to n.

Among the two proposed methods, sPCA generally achieves better performance than SDR. When the size of the true active set in the imputation model is small, that is, $c = 4$ (Table 2.1), all sPCA and SDR variations exhibit negligible bias and coverage near the nominal level. Within SDR methods, SDR_SIR tends to achieve slightly better performance than SDR_SAVE and SDR_PHD in terms of bias, MSE, and coverage. When the size of the true active set is large ($c = 100$), the improved performance of sPCA compared to SDR is more pronounced when correlation among the data is small ($\rho = 0.1$) or moderate ($\rho = 0.5$). As the number of variables in the true imputation model increases, the performance of the SDR methods slightly deteriorates. While SDR achieves satisfactory performance in terms of bias and MSE when c is small ($c = 4$), it exhibits modest bias when $c = 100$ and the correlation is small ($\rho = 0.1$ or $\rho = 0.5$), whereas the performance of sPCA methods remains satisfactory. The sPCA methods capture the maximum variance in the data and ultimately yields more favorable results. Although, SDR has modest bias for the case when $c = 100$, it outperforms the existing methods which exhibit substantial bias. The correlation among the data appears to have no effect on the performance of SDR. However, when $c = 100$, the sPCA methods have improved performance with increasing correlation. This suggest that when variables are strongly correlated, the sPCA methods provide sufficient information for imputation even though the variables screened in SIS may not be identical to the variables in the imputation model. As the dimension of data increases from $p = 200$ to $p = 1000$, the results are comparable for all values of p, c, and ρ. This result suggest that our methods can accommodate different size imputation models. In the case of $c = 100$, both our sPCA and SDR proposed methods have improved performance when the covariance structure is block diagonal with compound symmetric blocks (Table 2.4) as compared to the setting where the covariance matrix is first-order autoregressive (Table 2.2). Although SDR_SIR has superior performance within the SDR methods, it is important to note that among the sPCA methods, no method is preferred over the other.

Within the existing methods, in the case where $c = 4$, Blasso has better performance in terms of bias, MSE, and coverage rates except in the case $p = 200$ and $\rho = 0.9$. In that exception, DAlasso has better performance. Yet all existing methods underperform our proposed methods when $c = 100$. The KNN and RF methods exhibit extreme bias in all scenarios with the exception of KNN_S with $c = 100$, $p = 200$, $\rho = 0.5$, and block diagonal covariance structure. In addition, correlation amongst the data and the number of predictors also appear to have very little influence on the results of existing methods. In general, the existing methods do not perform well and our proposed approach using dimension reduction techniques yield more favorable results.

2.4 Data example

We apply the proposed methodology to a prostate cancer study (GEO GDS3289). The data set contains 104 samples of which 34 are benign epithelium samples and 70 nonbenign samples. There were 1,894 fully observed variables which were all used for screening in SIS. In this analysis, we are interested in conducting a logistic regression where we have a binary outcome (y) which is 1 if a sample is benign and 0 otherwise. The goal is to test whether a genomic biomarker VPS36 is associated with the outcome. However, VPS36 has 51% of its values subjected to missingness. For illustration purposes, we also include two fully observed biomarkers as predictors in the logistic regression model. In this analysis, the GS method is not applicable since we do not the underlying true data. In addition, the **mice** package used to conduct MI in R gives error messages, therefore are not included in our results.

In Table 2.5, we present the results of our analysis for estimating the parameter θ_1, that is the regression parameter associated with VPS36. There were 11 variables screened by SIS for VPS36. The sPCA, SDR, and existing methods give differing results in terms of estimates and p-value. For example, VPS36 is statistically significant using the sPCA methods for dimension reduction before MI but not in SDR. Yet, the direction of the point estimates are the same for all methods except RF. However, RF had extreme bias in the simulations and may be questionable in this application. In addition, the magnitude of the estimates for the sPCA methods are considerably larger but consistent across the four sPCA methods. In contrast, the parameter estimates and p-values are more variable in the three SDR methods. For example, the regression coefficients for VPS36 in the SDR method range from 0.517 to 1.260, with p-values of 0.542 and 0.190, respectively. Our simulations show that performing sPCA before MI generally yields minimal bias and adequate coverage rates, therefore, may be more preferred over the other methods in this application.

2.5 Discussion

Our work demonstrates the value of dimension reduction techniques in constructing imputation models in the presence of high-dimensional data, particularly when the size of the true imputation model is large. In the settings considered, the proposed methods outperform the existing methods, irregardless of the size of the true imputation model, the number of variables in the data set, and the correlation among the data. In comparing sPCA and SDR to construct imputation models, the sPCA method outperformed SDR in terms of bias, MSE, and coverage rate. A data example using genomics data from a prostate cancer study is used to further illustrate the usefulness of our proposed method.

We have considered settings under the MAR assumption where a single variable has missing values. In practice, more than one variable of interest may have missing values. Future work can extend our methods to the setting of general missing data patterns with more than one variable missing. In addition, it is of interest to develop methods to handle missing data under the assumption of missing not a random in the presence of high-dimensional data.

Table 2.1: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 4, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9

		$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
	Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	0.001	0.086	0.088	0.008	0.944	0.000	0.081	0.080	0.006	0.962	-0.003	0.080	0.082	0.007	0.956
	CC	-0.172	0.103	0.107	0.041	0.610	-0.170	0.099	0.101	0.039	0.582	-0.145	0.099	0.103	0.032	0.696
	MI	-0.662	0.251	0.108	0.449	0.226	-0.784	0.190	0.089	0.623	0.026	-0.778	0.209	0.085	0.612	0.026
$p = 200$	sPCA_ST	-0.026	0.100	0.097	0.010	0.940	-0.030	0.093	0.088	0.009	0.962	-0.015	0.094	0.097	0.010	0.956
	sPCA_PMD	-0.028	0.104	0.100	0.011	0.948	-0.027	0.097	0.091	0.009	0.970	-0.021	0.096	0.099	0.010	0.946
	sPCA_L	-0.031	0.102	0.097	0.010	0.938	-0.035	0.095	0.088	0.009	0.958	-0.018	0.095	0.098	0.010	0.960
	sPCA_AL	-0.030	0.101	0.097	0.010	0.938	-0.033	0.094	0.088	0.009	0.958	-0.016	0.094	0.098	0.010	0.960
	SDR_SIR	-0.028	0.106	0.107	0.012	0.940	-0.028	0.097	0.096	0.010	0.950	-0.031	0.094	0.094	0.010	0.932
	SDR_SAVE	-0.065	0.103	0.093	0.013	0.936	-0.067	0.098	0.088	0.012	0.900	-0.051	0.097	0.090	0.011	0.928
	SDR_PHD	-0.060	0.104	0.097	0.013	0.936	-0.063	0.098	0.089	0.012	0.918	-0.048	0.097	0.091	0.011	0.930
	Blasso	-0.059	0.113	0.098	0.013	0.944	-0.062	0.103	0.092	0.012	0.946	-0.074	0.100	0.088	0.013	0.908
	DAlasso	-0.155	0.163	0.099	0.034	0.944	-0.105	0.149	0.102	0.021	0.976	-0.049	0.126	0.094	0.011	0.988
	KNN_S	-0.254	0.150	0.143	0.085	0.640	-0.281	0.146	0.141	0.099	0.524	-0.289	0.142	0.126	0.100	0.468
	KNN_V	-0.313	0.158	0.126	0.114	0.504	-0.375	0.155	0.121	0.155	0.288	-0.444	0.153	0.120	0.211	0.110
	RF	-0.320	0.176	0.119	0.116	0.624	-0.333	0.172	0.114	0.124	0.560	-0.305	0.160	0.107	0.104	0.560
$p = 1000$	sPCA_ST	-0.016	0.100	0.094	0.009	0.970	-0.028	0.095	0.090	0.009	0.970	-0.015	0.093	0.087	0.008	0.958
	sPCA_PMD	-0.020	0.101	0.094	0.009	0.966	-0.031	0.096	0.090	0.009	0.968	-0.017	0.094	0.088	0.008	0.956
	sPCA_L	-0.023	0.102	0.094	0.009	0.964	-0.034	0.096	0.090	0.009	0.968	-0.019	0.094	0.088	0.008	0.956
	sPCA_AL	-0.021	0.101	0.094	0.009	0.968	-0.032	0.096	0.090	0.009	0.966	-0.017	0.094	0.088	0.008	0.954
	SDR_SIR	-0.028	0.097	0.096	0.010	0.950	-0.023	0.102	0.099	0.010	0.944	-0.034	0.097	0.095	0.010	0.932
	SDR_SAVE	-0.067	0.098	0.088	0.012	0.900	-0.065	0.098	0.090	0.012	0.910	-0.046	0.098	0.091	0.010	0.932
	SDR_PHD	-0.063	0.098	0.089	0.012	0.918	-0.061	0.099	0.093	0.012	0.926	-0.043	0.098	0.093	0.011	0.942
	Blasso	-0.093	0.121	0.120	0.023	0.910	-0.070	0.108	0.091	0.014	0.938	-0.079	0.102	0.096	0.015	0.940
	DAlasso	-0.277	0.180	0.109	0.089	0.794	-0.250	0.176	0.107	0.074	0.846	-0.176	0.179	0.115	0.044	0.954
	KNN_S	-0.306	0.151	0.141	0.113	0.486	-0.368	0.149	0.134	0.154	0.286	-0.396	0.146	0.136	0.175	0.218
	KNN_V	-0.310	0.156	0.116	0.109	0.518	-0.386	0.155	0.124	0.164	0.274	-0.449	0.154	0.117	0.215	0.112
	RF	-0.374	0.180	0.124	0.155	0.480	-0.389	0.178	0.115	0.164	0.396	-0.396	0.173	0.117	0.171	0.364

Table 2.1a: Percent of explained variance for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9 where $k = 1$

	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
$p = 200$	sPCA_ST	0.464	0.519	0.634
	sPCA_PMD	0.511	0.554	0.646
$p = 1000$	sPCA_ST	0.463	0.517	0.620
	sPCA_PMD	0.509	0.552	0.637

Table 2.1bi: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 4, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9, where k chosen to explain at least 80% of variance and ridge parameter set to $10e^{-5}$

		$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
	Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	0.001	0.086	0.082	0.007	0.962	0.000	0.082	0.082	0.007	0.950	-0.003	0.079	0.083	0.007	0.938
	CC	-0.176	0.102	0.102	0.042	0.604	-0.167	0.100	0.101	0.038	0.650	-0.145	0.099	0.103	0.032	0.702
	MI	-0.662	0.251	0.108	0.449	0.226	-0.784	0.190	0.089	0.623	0.026	-0.778	0.209	0.085	0.612	0.026
$p = 200$	sPCA_ST	-0.036	0.103	0.095	0.010	0.950	-0.038	0.097	0.093	0.010	0.950	-0.020	0.089	0.091	0.009	0.948
	sPCA_PMD	-0.033	0.103	0.095	0.010	0.960	-0.032	0.097	0.094	0.010	0.954	-0.020	0.090	0.092	0.009	0.950
$p = 1000$	sPCA_ST	-0.052	0.116	0.106	0.014	0.958	-0.028	0.101	0.098	0.010	0.952	-0.024	0.089	0.087	0.008	0.944
	sPCA_PMD	-0.045	0.117	0.106	0.013	0.956	-0.020	0.102	0.100	0.010	0.954	-0.024	0.089	0.087	0.008	0.952

Table 2.1bii: Mean, minimum, and maximum number of k principal components to estimate $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 4, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9 where k chosen to explain at least 80% of variance and ridge parameter set to $10e^{-5}$

		$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
	Method	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$p = 200$	sPCA_ST	9.686	7.000	12.000	7.768	5.000	10.000	3.142	1.000	5.000
	sPCA_PMD	8.170	7.000	10.000	6.716	5.000	9.000	3.018	1.000	5.000
$p = 1000$	sPCA_ST	9.523	7.000	12.000	8.018	5.000	10.000	3.734	2.000	7.000
	sPCA_PMD	7.974	6.000	9.000	7.294	5.000	10.000	3.524	2.000	6.000

Table 2.2: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9

		$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
	Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	0.011	0.156	0.158	0.025	0.940	-0.010	0.136	0.139	0.019	0.946	0.001	0.086	0.090	0.008	0.948
	CC	-0.334	0.168	0.172	0.141	0.518	-0.322	0.150	0.158	0.128	0.414	-0.233	0.112	0.113	0.067	0.484
	MI	-0.800	0.209	0.094	0.649	0.038	-0.772	0.216	0.089	0.604	0.040	-0.672	0.265	0.099	0.461	0.232
$p = 200$	sPCA_ST	-0.098	0.212	0.203	0.050	0.964	-0.067	0.174	0.162	0.031	0.972	-0.013	0.108	0.111	0.013	0.962
	sPCA_PMD	-0.105	0.214	0.204	0.053	0.966	-0.075	0.177	0.164	0.033	0.972	-0.016	0.110	0.112	0.013	0.962
	sPCA_L	-0.111	0.217	0.203	0.053	0.962	-0.085	0.180	0.166	0.035	0.972	-0.022	0.112	0.114	0.013	0.964
	sPCA_AL	-0.108	0.216	0.203	0.053	0.966	-0.079	0.179	0.165	0.034	0.972	-0.017	0.110	0.113	0.013	0.966
	SDR_SIR	-0.311	0.229	0.232	0.151	0.748	-0.242	0.200	0.189	0.094	0.820	-0.062	0.117	0.118	0.018	0.928
	SDR_SAVE	-0.153	0.217	0.209	0.067	0.956	-0.128	0.186	0.162	0.043	0.962	-0.077	0.119	0.108	0.018	0.938
	SDR_PHD	-0.188	0.224	0.218	0.083	0.920	-0.151	0.192	0.171	0.052	0.946	-0.071	0.120	0.112	0.018	0.936
	Blasso	-0.450	0.238	0.143	0.223	0.580	-0.420	0.219	0.122	0.191	0.546	-0.094	0.129	0.109	0.021	0.928
	DAlasso	-0.466	0.253	0.146	0.239	0.608	-0.393	0.231	0.130	0.171	0.682	-0.063	0.139	0.116	0.017	0.970
	KNN_S	-0.294	0.222	0.216	0.133	0.750	-0.279	0.205	0.193	0.115	0.738	-0.035	0.154	0.133	0.019	0.958
	KNN_V	-0.289	0.233	0.189	0.119	0.810	-0.310	0.217	0.161	0.122	0.760	-0.277	0.182	0.127	0.093	0.754
	RF	-0.442	0.250	0.147	0.217	0.664	-0.434	0.232	0.136	0.207	0.566	-0.246	0.182	0.117	0.074	0.836
$p = 1000$	sPCA_ST	-0.080	0.210	0.195	0.044	0.982	-0.065	0.175	0.165	0.031	0.978	-0.022	0.108	0.104	0.011	0.964
	sPCA_PMD	-0.085	0.212	0.195	0.045	0.978	-0.070	0.177	0.168	0.033	0.976	-0.025	0.110	0.105	0.012	0.964
	sPCA_L	-0.099	0.217	0.193	0.047	0.974	-0.084	0.182	0.168	0.035	0.980	-0.032	0.112	0.106	0.012	0.964
	sPCA_AL	-0.094	0.215	0.194	0.046	0.974	-0.079	0.180	0.168	0.034	0.980	-0.027	0.110	0.105	0.012	0.964
	SDR_SIR	-0.350	0.229	0.217	0.170	0.688	-0.314	0.209	0.201	0.139	0.702	-0.079	0.124	0.116	0.020	0.930
	SDR_SAVE	-0.135	0.215	0.192	0.055	0.956	-0.122	0.187	0.173	0.045	0.948	-0.078	0.119	0.105	0.017	0.944
	SDR_PHD	-0.183	0.225	0.199	0.073	0.926	-0.158	0.198	0.182	0.058	0.926	-0.076	0.122	0.106	0.017	0.952
	Blasso	-0.483	0.245	0.141	0.253	0.542	-0.483	0.232	0.131	0.250	0.456	-0.213	0.158	0.102	0.056	0.844
	DAlasso	-0.447	0.257	0.148	0.222	0.670	-0.408	0.240	0.139	0.185	0.700	-0.167	0.185	0.121	0.042	0.954
	KNN_S	-0.339	0.222	0.204	0.156	0.712	-0.320	0.207	0.195	0.140	0.688	-0.185	0.165	0.142	0.055	0.830
	KNN_V	-0.303	0.231	0.177	0.123	0.812	-0.302	0.217	0.169	0.119	0.770	-0.295	0.181	0.124	0.102	0.664
	RF	-0.470	0.250	0.136	0.240	0.608	-0.478	0.234	0.141	0.249	0.470	-0.357	0.195	0.111	0.140	0.636

Table 2.3a: Percent of explained variance for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9 where $k = 1$ with ridge parameter set to 0.00001

	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
$p = 200$	sPCA_ST	0.309	0.331	0.530
	sPCA_PMD	0.335	0.354	0.533
$p = 1000$	sPCA_ST	0.308	0.333	0.516
	sPCA_PMD	0.336	0.358	0.523

Table 2.3: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 4, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9

		$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
	Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	-0.004	0.066	0.066	0.004	0.956	0.003	0.061	0.062	0.004	0.942	0.003	0.059	0.058	0.003	0.948
	CC	-0.129	0.086	0.090	0.025	0.686	-0.105	0.082	0.079	0.017	0.756	-0.090	0.080	0.084	0.015	0.790
	MI	-0.742	0.205	0.102	0.561	0.070	-0.670	0.248	0.098	0.458	0.182	-0.642	0.278	0.089	0.421	0.326
$p = 200$	sPCA_ST	-0.034	0.088	0.084	0.008	0.954	-0.014	0.087	0.081	0.007	0.966	0.002	0.083	0.080	0.006	0.970
	sPCA_PMD	-0.029	0.092	0.086	0.008	0.958	-0.016	0.087	0.079	0.007	0.966	-0.004	0.080	0.078	0.006	0.958
	sPCA_L	-0.037	0.089	0.084	0.008	0.956	-0.015	0.087	0.081	0.007	0.966	-0.003	0.086	0.082	0.007	0.968
	sPCA_AL	-0.036	0.089	0.084	0.008	0.956	-0.014	0.087	0.081	0.007	0.966	0.001	0.084	0.081	0.006	0.972
	SDR_SIR	0.002	0.081	0.075	0.006	0.972	-0.015	0.074	0.073	0.006	0.954	-0.006	0.082	0.086	0.007	0.956
	SDR_SAVE	-0.059	0.091	0.076	0.009	0.940	-0.043	0.087	0.077	0.008	0.956	-0.073	0.092	0.086	0.013	0.902
	SDR_PHD	-0.052	0.091	0.076	0.009	0.944	-0.041	0.086	0.075	0.007	0.958	-0.066	0.091	0.086	0.012	0.926
	Blasso	0.042	0.096	0.089	0.010	0.948	0.040	0.091	0.091	0.010	0.942	0.028	0.091	0.083	0.008	0.968
	DAlasso	-0.185	0.169	0.098	0.044	0.938	-0.108	0.147	0.096	0.021	0.980	-0.005	0.109	0.080	0.006	0.992
	KNN_S	-0.187	0.137	0.126	0.051	0.742	-0.101	0.127	0.110	0.022	0.908	-0.055	0.119	0.105	0.014	0.952
	KNN_V	-0.256	0.147	0.109	0.077	0.644	-0.307	0.145	0.116	0.108	0.454	-0.375	0.144	0.124	0.156	0.234
	RF	-0.288	0.163	0.106	0.094	0.666	-0.422	0.148	0.096	0.187	0.126	-0.145	0.128	0.087	0.028	0.896
$p = 1000$	sPCA_ST	0.015	0.082	0.084	0.007	0.946	0.012	0.083	0.084	0.007	0.960	0.019	0.119	0.119	0.015	0.952
	sPCA_PMD	0.013	0.082	0.083	0.007	0.952	0.003	0.087	0.086	0.007	0.956	-0.002	0.120	0.118	0.014	0.946
	sPCA_L	0.013	0.083	0.083	0.007	0.946	0.013	0.083	0.084	0.007	0.958	0.015	0.118	0.118	0.014	0.950
	sPCA_AL	0.014	0.082	0.083	0.007	0.946	0.013	0.083	0.084	0.007	0.958	0.018	0.119	0.119	0.014	0.952
	SDR_SIR	0.001	0.081	0.083	0.007	0.944	-0.019	0.085	0.086	0.008	0.934	-0.116	0.141	0.157	0.038	0.874
	SDR_SAVE	-0.062	0.087	0.079	0.010	0.924	-0.029	0.087	0.080	0.007	0.948	-0.093	0.118	0.109	0.021	0.898
	SDR_PHD	-0.057	0.087	0.081	0.010	0.934	-0.026	0.087	0.081	0.007	0.954	-0.091	0.120	0.111	0.021	0.904
	Blasso	0.031	0.095	0.091	0.009	0.960	-0.118	0.159	0.196	0.052	0.926	-0.735	0.135	0.167	0.567	0.056
	DAlasso	-0.346	0.171	0.087	0.127	0.506	-0.433	0.168	0.102	0.198	0.200	-0.528	0.182	0.106	0.290	0.098
	KNN_S	-0.242	0.137	0.124	0.074	0.602	-0.359	0.129	0.122	0.143	0.174	-0.819	0.110	0.105	0.681	0.000
	KNN_V	-0.298	0.147	0.112	0.102	0.466	-0.616	0.131	0.118	0.394	0.004	-0.953	0.096	0.092	0.916	0.000
	RF	-0.356	0.167	0.105	0.138	0.422	-0.235	0.151	0.101	0.065	0.748	-0.719	0.129	0.102	0.528	0.000

Table 2.3bi: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0, 0.5, and 0.9, where k chosen to explain at least 60% of variance with ridge parameter λ varying as $1e^{-5}$, $1e^{-3}$, and $1e^{-1}$

			$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
		Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
		GS	0.011	0.155	0.154	0.024	0.960	-0.012	0.136	0.138	0.019	0.948	-0.003	0.085	0.084	0.007	0.960
		CC	-0.315	0.167	0.163	0.126	0.520	-0.329	0.150	0.158	0.133	0.418	-0.235	0.111	0.112	0.068	0.462
		MI	-0.800	0.209	0.094	0.649	0.038	-0.772	0.216	0.089	0.604	0.040	-0.672	0.265	0.099	0.461	0.232
$\lambda = 1e^{-5}$	$p = 200$	sPCA_ST	-0.186	0.223	0.209	0.078	0.910	-0.150	0.193	0.186	0.057	0.914	-0.023	0.109	0.103	0.011	0.952
		sPCA_PMD	-0.181	0.223	0.206	0.075	0.910	-0.146	0.193	0.188	0.057	0.918	-0.023	0.109	0.103	0.011	0.952
	$p = 1000$	sPCA_ST	-0.244	0.235	0.217	0.107	0.884	-0.211	0.208	0.200	0.085	0.884	-0.031	0.111	0.105	0.012	0.966
		sPCA_PMD	-0.243	0.236	0.218	0.106	0.882	-0.201	0.208	0.199	0.080	0.894	-0.030	0.111	0.106	0.012	0.970
$\lambda = 1e^{-3}$	$p = 200$	sPCA_ST	-0.187	0.223	0.209	0.078	0.910	-0.150	0.193	0.186	0.057	0.914	-0.022	0.109	0.104	0.011	0.952
		sPCA_PMD	-0.181	0.224	0.206	0.075	0.910	-0.147	0.193	0.188	0.057	0.918	-0.023	0.109	0.103	0.011	0.952
	$p = 1000$	sPCA_ST	-0.244	0.235	0.217	0.106	0.884	-0.211	0.208	0.200	0.084	0.884	-0.031	0.111	0.106	0.012	0.966
		sPCA_PMD	-0.243	0.236	0.218	0.106	0.882	-0.201	0.208	0.198	0.080	0.894	-0.030	0.111	0.106	0.012	0.970
$\lambda = 1e^{-1}$	$p = 200$	sPCA_ST	-0.213	0.225	0.189	0.081	0.910	-0.180	0.200	0.174	0.063	0.904	-0.027	0.125	0.113	0.013	0.966
		sPCA_PMD	-0.209	0.227	0.189	0.079	0.900	-0.179	0.201	0.176	0.063	0.902	-0.026	0.126	0.114	0.014	0.968
	$p = 1000$	sPCA_ST	-0.243	0.231	0.191	0.095	0.872	-0.217	0.207	0.179	0.079	0.888	-0.037	0.127	0.116	0.015	0.966
		sPCA_PMD	-0.247	0.233	0.191	0.098	0.884	-0.215	0.209	0.179	0.078	0.896	-0.036	0.128	0.117	0.015	0.964

Table 2.3bii: Mean, minimum, and maximum number of k principal components to estimate $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, q = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9 where k chosen to explain at least 60% of variance

		$\rho = 0.0$			$\rho = 0.5$			$\rho = 0.9$		
	Method	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$p = 200$	ST_sPCA	5.960	4.000	8.000	5.174	4.000	7.000	2.000	1.000	3.000
	PMD_sPCA	5.318	3.000	7.000	4.670	3.000	6.000	1.968	1.000	3.000
$p = 1000$	ST_sPCA	5.896	4.000	7.000	5.400	4.000	7.000	2.162	1.000	3.000
	PMD_sPCA	5.262	4.000	7.000	4.810	4.000	6.000	2.090	1.000	3.000

Table 2.3ci: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, q = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0, 0.5, and 0.9, where k chosen to explain at least 80% of variance with ridge parameter λ varying as $1e^{-5}$, $1e^{-3}$, and $1e^{-1}$

			$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
		Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
		GS	-0.001	0.153	0.151	0.023	0.964	0.002	0.136	0.137	0.019	0.950	-0.001	0.085	0.090	0.008	0.918
		CC	-0.333	0.165	0.175	0.142	0.498	-0.309	0.151	0.156	0.120	0.470	-0.232	0.111	0.119	0.068	0.458
		MI	-0.800	0.209	0.094	0.649	0.038	-0.772	0.216	0.089	0.604	0.040	-0.672	0.265	0.099	0.461	0.232
$\lambda = 1e^{-5}$	$p = 200$	sPCA_ST	-0.269	0.229	0.224	0.122	0.816	-0.189	0.200	0.190	0.072	0.894	-0.021	0.108	0.108	0.012	0.948
		sPCA_PMD	-0.250	0.226	0.224	0.113	0.846	-0.179	0.200	0.188	0.067	0.910	-0.020	0.109	0.108	0.012	0.944
	$p = 1000$	sPCA_ST	-0.317	0.237	0.225	0.151	0.780	-0.287	0.216	0.208	0.126	0.776	-0.032	0.114	0.104	0.012	0.964
		sPCA_PMD	-0.301	0.237	0.225	0.142	0.790	-0.274	0.215	0.207	0.118	0.806	-0.030	0.114	0.104	0.012	0.970
$\lambda = 1e^{-3}$	$p = 200$	sPCA_ST	-0.268	0.229	0.224	0.122	0.816	-0.189	0.200	0.189	0.072	0.892	-0.021	0.109	0.108	0.012	0.950
		sPCA_PMD	-0.250	0.226	0.224	0.113	0.846	-0.179	0.200	0.187	0.067	0.910	-0.020	0.109	0.108	0.012	0.946
	$p = 1000$	sPCA_ST	-0.317	0.237	0.224	0.150	0.780	-0.286	0.216	0.208	0.125	0.774	-0.031	0.114	0.105	0.012	0.964
		sPCA_PMD	-0.301	0.237	0.225	0.141	0.790	-0.273	0.215	0.206	0.117	0.806	-0.030	0.114	0.104	0.012	0.970
$\lambda = 1e^{-1}$	$p = 200$	sPCA_ST	-0.267	0.226	0.198	0.110	0.836	-0.191	0.200	0.173	0.066	0.904	-0.017	0.120	0.114	0.013	0.956
		sPCA_PMD	-0.258	0.225	0.199	0.106	0.862	-0.194	0.202	0.173	0.067	0.908	-0.018	0.124	0.118	0.014	0.958
	$p = 1000$	sPCA_ST	-0.282	0.230	0.195	0.118	0.832	-0.258	0.209	0.178	0.098	0.824	-0.033	0.125	0.110	0.013	0.972
		sPCA_PMD	-0.282	0.233	0.193	0.117	0.814	-0.260	0.211	0.179	0.100	0.822	-0.035	0.129	0.112	0.014	0.974

Table 2.3cii: Mean, minimum, and maximum number of k principal components to estimate $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, q = 100, p = 200$ or $p = 1000$ with autoregressive covariance matrix with ρ varying as 0.1, 0.5, and 0.9 where k chosen to explain at least 80% of variance

		$\rho = 0.0$			$\rho = 0.5$			$\rho = 0.9$		
	Method	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$p = 200$	sPCA_ST	11.190	8.000	14.000	10.028	8.000	13.000	4.458	3.000	7.000
	sPCA_PMD	9.672	8.000	11.000	8.752	7.000	10.000	4.090	3.000	6.000
$p = 1000$	sPCA_ST	11.094	9.000	14.000	10.394	8.000	13.000	4.916	3.000	8.000
	sPCA_PMD	9.518	8.000	11.000	8.982	7.000	11.000	4.548	3.000	7.000

Table 2.4: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9

		$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
	Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
	GS	0.001	0.107	0.115	0.013	0.954	0.002	0.073	0.074	0.005	0.954	-0.004	0.064	0.063	0.004	0.954
	CC	-0.227	0.131	0.133	0.069	0.596	-0.150	0.101	0.101	0.033	0.724	-0.115	0.092	0.094	0.022	0.772
	MI	-0.778	0.192	0.092	0.613	0.026	-0.502	0.301	0.109	0.264	0.770	-0.585	0.300	0.091	0.351	0.564
$p = 200$	sPCA_ST	-0.047	0.136	0.130	0.019	0.974	-0.002	0.100	0.097	0.009	0.950	-0.006	0.096	0.095	0.009	0.966
	sPCA_PMD	-0.052	0.138	0.130	0.020	0.968	-0.009	0.102	0.097	0.009	0.950	-0.007	0.097	0.095	0.009	0.964
	sPCA_L	-0.063	0.141	0.132	0.021	0.966	-0.009	0.102	0.098	0.010	0.952	-0.010	0.098	0.095	0.009	0.966
	sPCA_AL	-0.055	0.139	0.130	0.020	0.968	-0.005	0.101	0.098	0.010	0.950	-0.007	0.097	0.095	0.009	0.966
	SDR_SIR	-0.171	0.164	0.147	0.051	0.852	-0.027	0.107	0.106	0.012	0.940	-0.016	0.088	0.098	0.010	0.938
	SDR_SAVE	-0.123	0.151	0.120	0.030	0.942	-0.060	0.107	0.097	0.013	0.928	-0.027	0.094	0.093	0.009	0.954
	SDR_PHD	-0.138	0.157	0.129	0.035	0.924	-0.056	0.109	0.099	0.013	0.940	-0.026	0.093	0.093	0.009	0.952
	Blasso	-0.251	0.187	0.128	0.079	0.830	-0.020	0.125	0.103	0.011	0.988	0.035	0.101	0.089	0.009	0.966
	DAlasso	-0.284	0.200	0.123	0.096	0.826	-0.066	0.144	0.107	0.016	0.990	-0.039	0.133	0.097	0.011	0.990
	KNN_S	-0.122	0.174	0.155	0.039	0.930	0.008	0.138	0.117	0.014	0.982	-0.051	0.128	0.112	0.015	0.962
	KNN_V	-0.232	0.191	0.153	0.077	0.806	-0.306	0.169	0.127	0.110	0.596	-0.401	0.157	0.135	0.179	0.248
	RF	-0.370	0.204	0.130	0.154	0.622	-0.269	0.170	0.099	0.082	0.758	-0.242	0.153	0.102	0.069	0.714
$p = 1000$	sPCA_ST	0.010	0.099	0.106	0.011	0.934	0.055	0.074	0.079	0.009	0.870	0.014	0.102	0.096	0.009	0.954
	sPCA_PMD	0.004	0.100	0.106	0.011	0.936	0.044	0.076	0.080	0.008	0.906	-0.007	0.102	0.095	0.009	0.974
	sPCA_L	0.000	0.101	0.106	0.011	0.944	0.053	0.074	0.080	0.009	0.882	0.010	0.102	0.096	0.009	0.956
	sPCA_AL	0.006	0.100	0.106	0.011	0.934	0.055	0.074	0.080	0.009	0.872	0.013	0.102	0.096	0.009	0.956
	SDR_SIR	-0.068	0.130	0.125	0.020	0.934	0.001	0.084	0.085	0.007	0.936	-0.083	0.121	0.130	0.024	0.898
	SDR_SAVE	-0.076	0.114	0.102	0.016	0.936	-0.024	0.085	0.079	0.007	0.968	-0.095	0.108	0.099	0.019	0.902
	SDR_PHD	-0.075	0.119	0.109	0.017	0.944	-0.020	0.085	0.079	0.007	0.966	-0.095	0.110	0.101	0.019	0.908
	Blasso	-0.170	0.164	0.104	0.040	0.920	-0.086	0.139	0.146	0.029	0.972	-0.623	0.139	0.218	0.435	0.148
	DAlasso	-0.192	0.195	0.118	0.051	0.948	-0.209	0.183	0.141	0.063	0.922	-0.476	0.209	0.149	0.249	0.414
	KNN_S	-0.028	0.147	0.129	0.017	0.974	-0.206	0.130	0.114	0.055	0.654	-0.832	0.101	0.103	0.703	0.000
	KNN_V	-0.240	0.172	0.126	0.073	0.778	-0.575	0.140	0.120	0.345	0.004	-0.954	0.091	0.091	0.919	0.000
	RF	-0.347	0.186	0.109	0.132	0.616	-0.342	0.153	0.097	0.126	0.396	-0.738	0.116	0.092	0.553	0.000

Table 2.4a: Percent of explained variance for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9 where $k = 1$ with ridge parameter set to 0.00001

	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
$p = 200$	ST_sPCA	0.321	0.561	0.701
	sPCA_PMD	0.331	0.558	0.699
$p = 1000$	ST_sPCA	0.437	0.825	0.972
	sPCA_PMD	0.440	0.801	0.940

Table 2.4bi: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9, where k chosen to explain at least 60% of variance with ridge parameter λ varying as $1e^{-5}$, $1e^{-3}$, and $1e^{-1}$

			$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
		Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
		GS	0.000	0.107	0.108	0.012	0.960	0.002	0.072	0.071	0.005	0.954	-0.005	0.064	0.063	0.004	0.952
		CC	-0.233	0.131	0.141	0.074	0.566	-0.146	0.100	0.102	0.032	0.700	-0.113	0.092	0.092	0.021	0.784
		MI	-0.778	0.192	0.092	0.613	0.026	-0.502	0.301	0.109	0.264	0.770	-0.585	0.300	0.091	0.351	0.564
$\lambda = 1e^{-5}$	$p = 200$	sPCA_ST	-0.117	0.158	0.148	0.036	0.914	0.005	0.097	0.093	0.009	0.974	-0.004	0.094	0.095	0.009	0.962
		sPCA_PMD	-0.115	0.158	0.148	0.035	0.916	0.005	0.097	0.093	0.009	0.974	-0.004	0.094	0.095	0.009	0.962
	$p = 1000$	sPCA_ST	-0.006	0.117	0.112	0.013	0.962	0.055	0.074	0.079	0.009	0.870	0.014	0.102	0.096	0.009	0.954
		sPCA_PMD	-0.007	0.118	0.111	0.012	0.962	0.055	0.074	0.079	0.009	0.874	0.013	0.102	0.096	0.009	0.958
$\lambda = 1e^{-3}$	$p = 200$	sPCA_ST	-0.117	0.158	0.148	0.036	0.914	0.005	0.098	0.093	0.009	0.974	-0.003	0.094	0.095	0.009	0.964
		sPCA_PMD	-0.116	0.158	0.148	0.035	0.916	0.006	0.097	0.093	0.009	0.974	-0.004	0.095	0.095	0.009	0.962
	$p = 1000$	sPCA_ST	-0.006	0.117	0.112	0.013	0.962	0.056	0.074	0.080	0.009	0.870	0.014	0.102	0.096	0.009	0.954
		sPCA_PMD	-0.007	0.118	0.111	0.012	0.962	0.056	0.074	0.080	0.009	0.870	0.014	0.102	0.096	0.009	0.956
$\lambda = 1e^{-1}$	$p = 200$	sPCA_ST	-0.127	0.163	0.141	0.036	0.910	0.010	0.112	0.101	0.010	0.976	0.000	0.111	0.106	0.011	0.972
		sPCA_PMD	-0.132	0.166	0.143	0.038	0.910	0.011	0.113	0.102	0.010	0.976	-0.001	0.111	0.106	0.011	0.970
	$p = 1000$	sPCA_ST	-0.006	0.123	0.112	0.013	0.968	0.072	0.091	0.091	0.013	0.892	-0.150	0.131	0.103	0.033	0.856
		sPCA_PMD	-0.012	0.128	0.114	0.013	0.972	0.072	0.091	0.091	0.013	0.892	-0.150	0.131	0.103	0.033	0.854

Table 2.4bii: Mean, minimum, and maximum number of k principal components to estimate $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9 where k chosen to explain at least 60% of variance

		$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
	Method	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$p = 200$	sPCA_ST	5.420	4.000	7.000	1.782	1.000	3.000	1.048	1.000	2.000
	sPCA_PMD	5.062	4.000	6.000	1.780	1.000	3.000	1.042	1.000	2.000
$p = 1000$	sPCA_ST	3.984	2.000	6.000	1.000	1.000	1.000	1.000	1.000	1.000
	sPCA_PMD	3.694	2.000	5.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.4ci: Simulation results for estimating $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9, where k chosen to explain at least 80% of variance with ridge parameter λ varying as $1e^{-5}$, $1e^{-3}$, and $1e^{-1}$

			$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
		Method	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
		GS	0.008	0.107	0.105	0.011	0.966	0.001	0.073	0.074	0.005	0.944	0.004	0.065	0.065	0.004	0.950
		CC	-0.223	0.131	0.133	0.067	0.582	-0.142	0.101	0.109	0.032	0.696	-0.108	0.091	0.096	0.021	0.782
		MI	-0.778	0.192	0.092	0.613	0.026	-0.502	0.301	0.109	0.264	0.770	-0.585	0.300	0.091	0.351	0.564
$\lambda = 1e^{-5}$	$p = 200$	sPCA_ST	-0.141	0.162	0.152	0.043	0.898	-0.001	0.101	0.105	0.011	0.954	0.004	0.087	0.085	0.007	0.968
		sPCA_PMD	-0.137	0.162	0.152	0.042	0.910	0.001	0.100	0.104	0.011	0.956	0.005	0.087	0.085	0.007	0.970
	$p = 1000$	sPCA_ST	-0.051	0.126	0.112	0.015	0.968	0.054	0.075	0.074	0.008	0.904	0.014	0.102	0.096	0.009	0.954
		sPCA_PMD	-0.044	0.125	0.110	0.014	0.968	0.053	0.076	0.075	0.008	0.904	0.013	0.102	0.096	0.009	0.958
$\lambda = 1e^{-3}$	$p = 200$	sPCA_ST	-0.141	0.162	0.152	0.043	0.898	0.000	0.101	0.104	0.011	0.956	0.005	0.087	0.085	0.007	0.968
		sPCA_PMD	-0.137	0.162	0.151	0.042	0.910	0.001	0.101	0.104	0.011	0.956	0.006	0.087	0.085	0.007	0.970
	$p = 1000$	sPCA_ST	-0.050	0.126	0.112	0.015	0.968	0.056	0.075	0.074	0.009	0.904	0.014	0.102	0.096	0.009	0.954
		sPCA_PMD	-0.044	0.125	0.110	0.014	0.968	0.054	0.076	0.075	0.009	0.902	0.014	0.102	0.096	0.009	0.956
$\lambda = 1e^{-1}$	$p = 200$	sPCA_ST	-0.135	0.163	0.141	0.038	0.906	0.017	0.107	0.106	0.011	0.966	0.017	0.101	0.096	0.009	0.966
		sPCA_PMD	-0.145	0.168	0.143	0.041	0.904	0.004	0.113	0.112	0.012	0.968	0.021	0.102	0.097	0.010	0.966
	$p = 1000$	sPCA_ST	-0.039	0.126	0.107	0.013	0.972	0.071	0.092	0.085	0.012	0.906	-0.150	0.131	0.103	0.033	0.856
		sPCA_PMD	-0.047	0.132	0.110	0.014	0.966	0.067	0.093	0.088	0.012	0.914	-0.150	0.131	0.103	0.033	0.854

Table 2.4cii: Mean, minimum, and maximum number of k principal components to estimate $\hat{\theta}_1 = 1$ in the presence of missing data based on 500 Monte Carlo data sets where $n = 100, c = 100, p = 200$ or $p = 1000$ with block diagonal matrix with compound symmetric blocks and ρ varying as 0.1, 0.5, and 0.9 where k chosen to explain at least 80% of variance

		$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
	Method	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$p = 200$	sPCA_ST	10.384	8.000	13.000	4.924	2.000	10.000	2.004	1.000	3.000
	sPCA_PMD	9.108	8.000	11.000	4.258	2.000	6.000	2.010	1.000	3.000
$p = 1000$	sPCA_ST	8.988	6.000	12.000	1.210	1.000	3.000	1.000	1.000	1.000
	sPCA_PMD	7.964	6.000	10.000	1.246	1.000	3.000	1.000	1.000	1.000

Table 2.5: Estimation of the predictor variable θ_1 that is associated with the incomplete biomarker VPS36 in logistic regression using complete case analysis (CC), four sparse PCA methods (sPCA), three SDR methods, Bayesian Lasso (Blasso), direct use of adaptive lasso (DAlasso), and random forest multiple imputation (RF) using the prostate cancer study

Method	Estimate	SE	p-value
CC	1.336	1.165	0.2515
Blasso	0.226	0.692	0.7463
DAlasso	0.908	0.985	0.3631
RF	-2.922	2.002	0.1550
sPCA_ST	2.290	0.942	0.0183
sPCA_PMD	2.276	0.945	0.0192
sPCA_L	2.170	0.925	0.0228
sPCA_AL	2.256	0.940	0.0199
SDR_SIR	1.260	0.950	0.1898
SDR_SAVE	0.864	0.813	0.2928
SDR_PHD	0.517	0.842	0.5418

Chapter 3

Nonparametric Imputation for Nonignorable Missing Data

Abstract

Missing data is a very common phenomenon in studies across different disciplines. Multiply imputing the missing values with several plausible values accounts for the uncertainty about the underlying true values and is a popular technique due to its ease of use. However, the vast majority of imputation techniques are designed for an ignorable missing data mechanism since nonignorability is an assumption more challenging to handle. Under non-ignorable missingness, one assumes the nonresponse mechanism depends on unobserved values, and the outcome model for the variable with missing values and the nonresponse model must be modeled jointly. Consequently, joint modeling can produce results that are sensitive to the failure of the two working models. We propose a more robust, nonparametric technique to multiply impute missing data in the presence of non-ignorability by allowing users to choose optimal weights so the resulting estimator can rely more heavily on the working model that is more likely to be correctly specified. Using the two working models, we derive predictive scores to achieve dimension reduction and use the resulting scores coupled with a nearest neighbor hot deck to multiply impute the missing values. By adopting the predictive scoring technique, we allow the covariates to be either categorical or continuous. The new proposed method is shown to outperform several existing multiple imputation methods for non-ignorable missing data in simulations. In addition, the method is illustrated using a real data example from the Georgia Coverdell Acute Stroke Registry.

3.1 Introduction

The ultimate goal of all analyses is to make valid inference from the data. Yet missing data often occur which compromise data quality, and more importantly, introduce bias and loss of efficiency which can lead to invalid inference. Rubin (1987) created a taxonomy for missing data to accommodate its complexity which also determine the appropriateness of analysis methods. Data are missing completely at random (MCAR) if the probability of an observation being missing does not depend on observed or unobserved variables. The assumption of missing at random (MAR) is a less stringent assumption that occurs when the probability of missing data are related to some other completely measured variable(s) in the model. When neither MCAR nor MAR hold, then the data are missing not at random (MNAR). Test have been developed (Little, 1988b) to distinguish between MCAR and MAR, but it is difficult to distinguish between MAR and MNAR. If questions arise about the source of missingness, then a more realistic assumption is that the data are MNAR.

In the analysis of missing data, one must also consider whether the missing data mechanism is ignorable. Ignorability is a standard assumption but it is only appropriate when the data is MCAR or MAR and the parameters governing the missing data process are distinct from parameters to be estimated (Rubin, 1976). Although ignorability is a convenient assumption, it is usually unrealistic when missingness is not by design. Suppose, for example, some healthcare professionals are more reluctant to evaluate acuity of stroke patients if they assume the stroke is not severe. Stroke severity may depend on many characteristics, which may not all be measured therefore the nonresponse mechanism is not ignorable. Since a non-ignorable missing data mechanism depends on unobserved data, making inference while accounting for non-ignorability is typically not trivial. Generally such attempts are complicated by the need to simultaneously account for the propensity of missingness model and the outcome model. A common practice is to analyze nonignorable missing data through the use of the selection model (Heckman, 1979) and pattern-mixture models (Little, 1993). Selection and pattern mixture models augment the model for the complete data with a missing data model. Selection models allow for specification of the full-data distribution, and therefore require assumptions be made about the data distribution of the unobserved values. Alternatively, pattern-mixture models stratify the data by missing data patterns, thus the response process is a mixture of models of different missing data patterns. However, these methods are computationally complex and not implemented in most statistical software.

Imputation is a commonly used method for handling missing data due to its ease of use. Imputation methods include multiple imputation and bootstrap imputation. Multiple imputation (Rubin, 1987) replaces missing values with multiple "plausible" values drawn from the posterior predictive distribution to create M complete data sets. MI allows for standard completed-data analysis, therefore, has become a popular technique due to its ease of use. Bootstrap imputation (BI) (Efron, 1994) generates multiple bootstrap samples from the original unimputed sample with missing values and fills in the missing data in the bootstrap samples using imputation procedures. However, there has been limited development of the two latter methods to handle non-ignorable nonresponse. Despite the popularity and advancement of these methods, there has been limited development of these methods to handle the case of nonignorable nonresponse. Motivated by such facts, we seek to develop an imputation method to handle missing data that arise from non-ignorable missing data mechanisms.

Although limited imputation methods exist for non-ignorable missingness, there are some approaches to perform parametric imputation. Glynn et al. (1993) assumed a random fully observed follow-up sample was available and used pattern-mixture models assuming a different normal distribution for the respondents, subjects with observed values, and nonrespondents or subjects with missing values. The nonrespondent distribution was assumed to follow the same distribution of the nonrespondents that were eventually available for follow-up. Carpenter et al. (2007) proposed a reweighting approach which involved deriving weights using importance sampling and using the weights to adjust missing at random multiple imputation estimates to account for non-ignorability. Under the approach, Carpenter et al. (2007) derived an imputation estimate and variance estimator but the variance was underestimated. Similarly, Kim and Kim (2012) also adjust for non-ignorability by deriving importance sampling weights coupled with the Expectation-Maximization (EM) algorithm. However, Kim and Kim (2012) apply a fraction of the original weight of the nonrespondents such that the sum of the fractional weights from all the matched respondents is equal to the original weight of the nonrespondents. Alternatively, Siddique et al. (2012) developed an imputation procedure that incorporates missing data mechanism uncertainty by specifying a range of ignorability assumptions and combining these assumptions into one inference. Jolani (2012) suggested a random indicator (RI) method which involves iteratively drawing imputations from the incomplete variable and the response indicator to determine the difference of the adjustment of the observed values from the unobserved values. A more recent contribution to the development of parametric imputation for non-ignorable nonresponse was proposed by Sullivan and Andridge (2015). Sullivan and Andridge (2015) extended the pattern-mixture model and the hot deck (PMMHD) to handle non-ignorability by specifying a sensitivity parameter that determines the missingness mechanism. We compare the latter methods, namely the RI method and PMMHD, in simulation studies.

Nonparametric imputation methods are far less developed than parametric methods for non-ignorability. Glynn et al. (1993) first adjust the approximate Bayesian bootstrap (ABB) to accommodate non-ignorability by randomly drawing with replacement from followed-up nonrespondents and then drawing again with replacement from the same sample for the nonrespondents (not in the follow-up). Siddique and Belin (2008) extend the ABB but instead by using predictive mean matching in a hot deck. Bootstrap samples were drawn with probability proportional to a function of the observed values where the function is chosen by the imputer and dependent on whether the imputer believes the values are higher or lower for the nonrespondents. A regression on the bootstrap sample of observed values is used to create predicted values that are used in a distanced based hot deck. Our new nonparametric MI method also harnesses the power of bootstrapping and the hot deck but is distinctive from these methods.

We propose a nonparametric imputation method based on bootstrap imputation and multiple imputation for non-ignorable nonresponse using an iterative procedure. Suppose we have an incomplete variable Y that is partitioned $Y = (Y_{obs}, Y_{mis})$ into observed values, Y_{obs}, and missing values, Y_{mis}, with a response indicator R that identifies the pattern of missing data. We define two models, one for predicting the response indicator R and the other for predicting Y_{mis}. The working model for the response indicator R includes the incomplete variable Y, hence is non-ignorable. We use the two models to derive predictive scores that are standardized to stabilize the imputation. For each subject with missing values, we randomly draw from a neighborhood of observed values whose predicted scores are close to the predicted scores for the subjects with missing values as in a nearest neighbor hot deck (Sande, 1982). We iterate between the two models to obtain imputations for Y_{mis}. However, conducting MI does not ensure the imputation method is 'proper' in the sense of yielding valid inferences that reflect variability (Rubin, 1987). In order to yield proper imputations, a multiple imputation procedure must propagate imputation uncertainty. To overcome this issue, we also propose a new method that relies on the use of bootstrap imputation (Efron, 1994). Our methods have several advantages over the existing parametric techniques for nonignorable nonresponse. Unlike existing approaches for non-ignorable missing data, our new proposed approach allows a user to specify weights based on confidence in the outcome model and the nonresponse model which ultimately improves efficiency. Furthermore, our approach is more robust to misspecification because it allows the user to specify weights and does not use the models directly to impute the missing values, therefore is nonparametric. In addition, our methods do not rely on follow-up data to be obtained on the unobserved cases, is more informative in the sense that it does not produce a range of inferences, and does not rely on specification of a sensitivity parameter that determines the missingness mechanism or adjustment for missingness. In contrast to existing nonparametric approaches, our methods can alleviate the curse of dimensionality in the presence of numerous covariates in the two models by the specification of models that allow for dimension reduction.

The remainder of the chapter is organized as follows. In Section 3.2, we present the notation and methodology for our new approach. In Section 3.3, we present results of a simulation study to evaluate our method and compare it to the parametric methods of Jolani (2012) and Sullivan and Andridge (2015). In Section 3.4, we illustrate the new proposed method using a real-world data example with stroke registry data. In Section 3.5, we conclude with a brief discussion.

3.2 Methodology

Let Y denote a single variable with missing values where Y_{obs} denotes the observed components of Y and Y_{mis} denotes the missing components. We let n_{obs} be the number of observed cases which is the sample size of Y_{obs}. Similarly, we let n_{mis} be the number of unobserved cases which is the sample size of Y_{mis}. We also have a set of fully observed variables denoted $\boldsymbol{X}$ that are predictive of either the variable with missing values Y or the response indicator of Y. The response indicator for Y, denoted R is defined as 1 if Y is observed and 0 if Y is missing. The fully observed variables $\boldsymbol{X}$ can also be divided into two components, namely $\boldsymbol{X}_{mis}$ and $\boldsymbol{X}_{obs}$, which are subjects with missing outcome Y and observed Y, respectively. The collection of observed data is denoted by $\boldsymbol{O} = \{Y_{obs}, \boldsymbol{X}, R\}$.

Our proposed method extends the work of Long et al. (2012) to account for non-ignorable missingness by incorporating the variable with missing values Y in the propensity model for the response indicator R. We developed a new imputation procedure that iteratively imputes Y_{mis} using a nearest neighbor hot deck approach and fits two working models which include the propensity model for the missing data mechanism and the data model for Y.

In this section we present two methods for implementing nonparametric imputation for non-ignorable missing data which include the use of bootstrap imputation and multiple imputation.

3.2.1 Nonparametric Bootstrap Imputation

Under a non-ignorable missing data mechanism, parameters associated with the outcome model and the nonresponse model should be estimated simultaneously. In order to estimate these parameters we adopt the ideas of predictive mean matching (Little, 1988a) and specify a model for the outcome and response models. We use predictive mean matching to obtain standardized scores, which are then used to calculate the distances from donors to recipients. The distance is used to find a neighborhood that consists of K donors who have the smallest K distance from the recipients as in K nearest neighbors hot deck imputation (Sande, 1982).

The detailed steps of our bootstrap imputation procedure in the case of a univariate missing data pattern is as follows:

1. Bootstrap: To account for uncertainty in estimating the parameters in the postulated working models, we first generate a bootstrap sample of the complete cases and incomplete cases by selecting n samples with replacement. We denote these bootstrap observed samples as $\{Y_{obs}^{b_m}, \boldsymbol{X}_{obs}^{b_m}\}$, where b_m distinguishes the

bootstrap sample from the original sample and the subscript $m = 1, ..., M$ is the number of imputed data sets.

2. Initial values: Start with the initially completed-data $Y^{(0)} = \{Y_{mis}^{(0)}, Y_{obs}^{b_m}\}$. The initial values for Y_{mis} is made by a regression of $Y_{obs}^{b_m}$ on $\boldsymbol{X}_{obs}^{b_m}$ using the first n_{obs} bootstrapped complete cases. The estimates from the regression of the bootstrapped complete cases are used to make initial predictions of Y_{mis} denoted $Y_{mis}^{(0)}$. Other approaches for initializing $Y_{mis}^{(0)}$ such as randomly drawing $Y_{mis}^{(0)}$ from Y_{obs} are also appropriate.

3. Working models: Building from the work of Long et al. (2012), we have a working model for predicting Y_{mis} and for predicting the missing data indicator R. We define an iteration t where $t = 1, 2, ..., T$ and iteratively impute Y_{mis} by fitting the two working models and calculating predictive scores.

   (i) The first working model is for the outcome Y,

$$E(Y^{(t-1)}|\boldsymbol{X}_1) = l_1(\boldsymbol{X}_1; \boldsymbol{\beta}). \tag{3.1}$$

   In model (1) the function l_1 is a specified real-valued, smooth function, $\boldsymbol{X}_1$ is a set of p_1 fully observed predictors, and $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_{p_1})^T$ is a vector of regression coefficients. At iteration t, the model (3.1) is fit using $Y^{(t-1)}$ and $\boldsymbol{X}_1$ from all n observations. We use linear regression models for continuous Y, Poisson loglinear models for count Y, and multinomial logistic regression for Y with J categories. These models can include as many covariates as possible for $\boldsymbol{X}_1$.

   (ii) We have a second working model for predicting the probability of response,

$$E(R|\boldsymbol{X}_2, Y^{(t-1)}) = l_2(\boldsymbol{X}_2, Y^{(t-1)}; \boldsymbol{\alpha}) \tag{3.2}$$

where l_2 is a specified real-valued smooth function, $\boldsymbol{X}_2$ is a set of p_2 fully observed covariates, and $\boldsymbol{\alpha}$ is a vector of regression coefficients. The nonresponse process depends on the incomplete variable Y, thus the missingness mechanism is non-ignorable. It is important to note that the adjustment for non-ignorable missingness is made by including the incomplete variable in the model for the probability of missingness. Initial values are filled in for the missing values of Y but are then iteraively imputed to adjust for non-ignorable nonresponse. The clear distinction from the proposed method and the previous work of Long et al. (2012) is the inclusion of $Y = (Y_{mis}, Y_{obs})$ into the model for the missingness mechanism. In addition, dimension reduction models, such as the lasso or ridge regression, can be used to fit these models. Dimension reduction techniques alleviate the curse of dimensionality in the presence of numerous covariates in the two models.

4. Predictive scores: We obtain estimated predictive scores from the two working models (1) and (2) by

$$(P_1^{(t)}, P_2^{(t)}) = \{l_1(\boldsymbol{X}_1; \hat{\boldsymbol{\beta}}^{(t)}), l_2(\boldsymbol{X}_2, Y^{(t-1)}; \hat{\boldsymbol{\alpha}}^{(t)})\} \tag{3.3}$$

where $\hat{\boldsymbol{\alpha}}^{(t)}$ and $\hat{\boldsymbol{\beta}}^{(t)}$ are the estimated coefficients from the outcome model (1) and the nonresponse model (3.2), respectively. The predictive scores (P_1, P_2) are centered and scaled to have mean 0 and variance 1 to stabilize the imputation.

5. Distances: Given P_1 and P_2, for each subject with missing Y we create an imputing set that consists of observed responses from subjects who are similar.

For each i in the sample of subjects with missing values, the estimated standardized predictive scores $(P_1^{(t)}, P_2^{(t)})$ are used to define the distance between subjects $i = 1, ..., n_{mis}$ (nonrespondents) and $j = 1, ..., n_{obs}$ (respondents) as

$$d(i,j) = \{\omega_1[P_1^{(t)}(i) - P_1^{(t)}(j)]^2 + \omega_2[P_2^{(t)}(i) - P_2^{(t)}(j)]^2\}^{(1/2)} \tag{3.4}$$

where ω_1 and ω_2 are positive weights for the two predictive scores. We let ω_1 and ω_2 satisfy the condition that $\omega_1+\omega_2 = 1$ and choose the weights by the amount of confidence an imputer has on the specification of ω_1 and ω_2, for the outcome and response models, respectively.

6. Nearest neighbors: For each bootstrapped observation i with missing values, we find a set of similar observed subjects in the bootstrap sample by using the distance $d(i,j)$ to define the set of $K-$nearest neighbors, denoted by $R_K(i)$, that consists of K donors from $j = 1, ..., n_{obs}$ that have the smallest K distances from observation i. An update for Y_{mis} is created by randomly drawing from $R_K(i)$ with equal probability and we define the t^{th} iteration for Y as $Y^{(t)} = \{Y_{obs}^{b_m}, Y_{mis}^{(t)}\}$. The imputation is nonparametric because ultimately we use the nearest neighbor hot deck to make imputations.

7. We repeat steps 3 through 6 until convergence. The algorithm usually converges after a few iterations. After the algorithm converges, the last drawn $Y_{mis}^{(t)}$ will be one set of imputed values for Y_{mis}, denoted by $Y_{mis}^{imp,1,b_1}$ where superscript b_1 denotes that the missing values were imputed for the bootstrap sample of the missing values. One complete data set is created by using the bootstrapped n_{obs} observations and the imputed $Y_{mis}^{imp,1,b_1}$.

Multiple M bootstrap imputations can be generated by starting from different bootstrap samples in step 1 to create M complete data sets composed of $\{Y_{mis}^{imp,1,b_1}, Y_{obs}, \boldsymbol{X}\}$,

$\{Y_{mis}^{imp,2,b_2}, Y_{obs}, \boldsymbol{X}\}$, ...,$\{Y_{mis}^{imp,M,b_M}, Y_{obs}, \boldsymbol{X}\}$. Each data set completed by imputation can be analyzed using the standard complete-data methods.

We use combining rules for bootstrap imputation (Efron, 1994) to get a combined nonparametric bootstrap imputation (NBI) estimate we denote as $\bar{\gamma}_{NBI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\gamma}_m$ where the variance of $\bar{\gamma}_{NBI}$ is $\hat{V}_{BI} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\gamma}_m - \bar{\gamma}_{NBI})^2$.

3.2.2 Nonparametric Multiple Imputation

An alternative approach for imputation is multiple imputation proposed by Rubin (1976). We incorporate a bootstrap scheme which allows for resampling of only the observed data and accounts for the variability in estimating the parameters in the imputation. We create initial values as in step 1 in bootstrap imputation but only using the bootstrapped observed data. We fit the models (3.1) and (3.2) using the bootstrapped data. We compute the distance between subject i with missing outcome and all other subjects that have an observed outcome in the bootstrap sample. The algorithm continues as in bootstrap imputation but with the key distinction that the completed data sets are composed of the original observed data (pre-bootstrap) where $\{Y_{mis}^{imp,1}, Y_{obs}, \boldsymbol{X}\}$, $\{Y_{mis}^{imp,2}, Y_{obs}, \boldsymbol{X}\}$,..., $\{Y_{mis}^{imp,M}, Y_{obs}, \boldsymbol{X}\}$ are the completed M data sets. Standard complete data analysis are performed on each imputed data set. Similarly to bootstrap imputation, we let $\hat{\theta}_m$ and W_m, $m = 1, ..., M$ be the estimates and associated variances, respectively, obtained from each M analyses. We use Rubin's combining rules (Rubin, 1987) to get a combined nonparametric multiple imputation (NMI) estimate denoted $\bar{\theta}_{NMI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$ and the total variability associated with $\bar{\theta}_{NMI}$ is $T_{NMI} = \bar{W} + \frac{M+1}{M}B$ where $\bar{W} = \frac{1}{M}\sum_{m=1}^{M}W_m$ and $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2$ to create one inferential summary.

3.3 Simulation studies

In this section, we carry out simulation studies to evaluate the performance of our proposed methods in finite samples. In our simulations, we focus on estimating the mean in the presence of missing data with non-ignorable nonresponse. We are also interested in evaluating the impact of incorrect specification of the data and response models, correlation amongst the data, and the choices of the user-specified weights (ω_1, ω_2). For each simulation scenario our goal was to examine the following measures: the average relative bias (RB) computed using the ratio of the bias to the absolute value of the nonzero true value; the standard deviation (SD) of the resulting estimates; the average standard error (SE); the mean squared error (MSE); the coverage rates (CR) of 95% Wald confidence intervals. The following methods are compared: complete-case (CC) analysis which discards subject with missing observations; a parametric multiple imputation (PMI) method, where Bayesian linear regression with the fully observed covariates is fit and then used to draw imputations for each missing observation; the random indicator (RI) method (Jolani, 2012); the pattern-proxy mixture hot deck (PPMHD) proposed by Sullivan and Andridge (2015); our proposed nonparametric multiple imputation (NMI) and nonparametric bootstrap imputation (NBI) method. The PMI and PPMHD methods involve fitting a regression model for Y, and the RI, NBI, and NMI methods also involve fitting a regression model for the response indicator R. The models were fit using the method of maximum likelihood.

3.3.1 Setup

Two set of simulations were carried out with 500 replicates and sample size $n = 400$. Five fully observed covariates are generated from the multivariate normal distribution:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \right)$$

with 3 different correlations, $\rho = \{-0.5, 0.0, 0.5\}$. The hypothetical complete outcome of interest Y was generated from a normal distribution with mean $E(Y|\boldsymbol{X}) = 10 + X_1 + X_2 + X_3 + X_4 + X_5$ and variance of 9 to result in 10 for the true mean of Y. The outcome Y was subject to missingness with the probability of being missing generated from a logit model, $logit[P(R = 1|\boldsymbol{X})] = -4X_1 + 2X_2 + 2X_3 + 2X_4 - 2X_5 - 0.25Y$, where R is the response indicator which takes the value of 1 if Y is missing and 0 otherwise. The aforementioned model depends on the variable with missing values Y, hence the missing data mechanism is non-ignorable. Each Monte Carlo data set includes the variable with missing values Y and the fully observed variables $\boldsymbol{X}$. In the first simulation scenario, we estimate the mean of the incomplete variable Y with both models for the outcome Y and the nonresponse model correctly specified. In the second scenario, both models are incorrectly specified where the outcome Y was subject to missingness with the probability of being missing generated from a logit model, $logit[P(R = 1|\boldsymbol{X})] = -4X_1 + 2X_2 + 2X_3 + 2X_4 + 2X_5 + 0.2Y$ and the true model for the outcome is $Y = 8 + X_1 + X_2 + X_3 + X_4 + \epsilon$. However, the working model for the outcome only includes X_2, X_3, X_4 to create a misspecification and the model for the response includes Y in addition. We also consider misspecification of either the response or outcome models. The true outcome model remains the same as in the second scenario. However, in the third scenario the response model is incorrectly specified with the true response model a logistic regression model $\text{logit}\,[P(R = 1)] = -4X_1 + 2X_2 + 2X_3 + 2X_4 + 6X_5 + 0.2Y$ and a misspecification is created by only using X_1, X_2, X_3 and Y to fit the model. To misspecify the outcome model in the last scenario, we exclude X_4 from the model.

The NBI and NMI estimate of the mean is shown as NBI(K, ω_1, ω_2) and NMI(K, ω_1, ω_2) where K is the number of nearest neighbors used in the imputation. The weighting scheme was varied as $(\omega_1, \omega_2) = (0.2, 0.8), (0.5, 0.5)$, and $(0.8, 0.2)$. The specification of (ω_1, ω_2) provides a natural way to incorporate ones prior beliefs on the validity of the two working models. In simulations we chose $K = 3$ neighbors as recommended by Long et al. (2012). In simulations, Long et al. (2012) show increasing K beyond 3 leads to larger bias and slightly higher MSEs. We choose $M = 50$ bootstrap imputations for NBI, $M = 30$ multiple imputations for NMI, and used 100 iterations for each imputation. In the statistical literature, it has been shown these values of M achieve good performance in finite samples. The mean estimate using the PPM hot deck is shown as PPMHD(c, λ) where c is the 'closeness' parameter that determines the closeness of the donor and the recipient in distanced-based donor selection. The value of λ varies the missing data mechanism. We chose $c = 3$ in simulations as suggested in Sullivan and Andridge (2015) and take $\lambda = 0$ to assume MAR, $\lambda = 1$ to assume weak MNAR, and $\lambda = \infty$ to assume a strong MNAR mechanism.

3.3.2 Results

The simulation results for estimating the mean of the incomplete variable Y in the presence of non-ignorable nonresponse is presented in Table 3.1 through Table 3.4. The missing probabilities ranged from 32% to 35%. Note that the CC and PMI estimate of the mean suffered from notable undercoverage in both cases. In Table 3.1, the models for Y and R were correctly specified. When both models are correctly specified, the bias was negligible for all estimators except for the RI estimator. The RI estimator exhibited substantial bias. However, the NBI and NMI were least biased with coverage rates closer to the 95% nominal level. While the bias was negligible for all NBI and NMI estimators, the method lead to an even smaller bias when a larger weight was assigned to the nonresponse model for $R(\omega_2)$. As the weight ω_2 increases from 0.2 to 0.8 for the model for R, the bias of the NBI and NMI estimator decreased slightly. The impact of weights on SE is minimal for our estimator. The MSE of our proposed methods was lower as compared to the other methods. Our proposed method and the RI method have coverage near the nominal 95% confidence level. However, we note that coverage rates of the RI estimator can be misleading in this case, since it usually exhibits large sample variation in addition to the extreme bias. The NBI usually had slightly better coverage rates and smaller MSE as compared to our NMI estimator. In both scenarios, the PPMHD has less bias and better coverage when the closeness parameter is $c = 3$ and $\lambda = 1$ which is assuming a less extreme case of MNAR. Although the PPMHD had negligible bias, there was significant undercoverage. A plot of the standard errors for the 500 datasets showed that the standard errors were generally lower than our methods. However, large standard errors that were potentially outliers resulted in large estimates of the average SE. In both simulation scenarios, all methods perform better (in terms of bias and coverage) when the correlation among the data increases from -0.5 to 0.5.

In Table 3.2, we present the results with both models incorrectly specified. Even though the models are misspecified, the simulation results show that the NBI estimators of the mean are nearly unbiased when the correlation among the data is 0.0 and 0.5. An advantage of our approach is that the NBI and NMI method is nonparametric and only uses the predictive scores to evaluate the similarity between subjects, hence its dependence on the two models is weaker than that of the other methods. All other methods are either biased or exhibit undercoverage of the confidence interval or both. The RI method exhibits considerable overcoverage of the 95% confidence level due to inflated standard errors. In contrast, the PMI and PPMHD methods have substantial undercoverage with incorrect specification of both models. Although we noticed favorable results with the NBI method, using NMI to impute missing values and estimate the mean shows large bias, larger MSE, and considerable undercoverage. In Table 3.3 and 3.4, we present results of either the outcome or response model misspecified. However, our results are only a slight improvement over the existing methods in terms of bias, MSE, and coverage. When the correlation among the data is -0.5 and outcome model is misspecified, the RI method fails.

In summary, the proposed NBI estimator and NMI estimators achieve similar or better performance in all settings compared with other estimators considered in our simulation studies. Our results suggest that it is more important to correctly specify the model for the nonresponse R. Thus, it is recommended to choose larger w_2 values, possibly greater than or equal to 0.5, in the absence of prior knowledge on the models.

3.4 Motivating Example

We illustrated the proposed methods using the Georgia Coverdell Acute Stroke Registry (GCASR) data set. A detailed description of the registry data set was reported elsewhere (Camp et al., 2015). The primary goal of the GCASR program is to improve the care of acute stroke patients in the hospital and pre-hospital setting. We used a sample of 2,115 patients with acute ischemic stroke in which data were collected from multiple sources, including patient surveys and medical records, between 2007 and 2011. Fully observed variables included age, gender, race, and dichotomous variables for obesity, diabetes, atrial fibrillation, hypertension, IV TPA, and whether or not the patient had insurance. Other clinical variables collected for all patients were cholesterol level, systolic blood pressure, diastolic blood pressure,and the hospital. While there were many variables with missing values, for illustrating our method we focus on imputation of the National institute of Health stroke score (NIHSS). The

NIHSS is a clinical assessment tool to measure stroke severity, predict patient clinical outcome, and is often used to determine appropriate treatment. Of the 2115 patients, 681 were missing NIHSS (32% missing). Clinicians often believe the assessment is time-consuming and typically do not calculate the score if they do not believe the patient might be a candidate for acute treatment with intravenous thrombolytics or enrollment in investigational protocols (Richardson et al., 2006). Furthermore, these barriers in implementation of the score created concern for whether or not the data were nonignorably missing. To account for a potential non-ignorable missing data mechanism, we applied our new proposed method to estimating the overall mean NIHSS in the study population.

We imputed the NIHSS using the NMI and NBI approach of Section 3.2 by first constructing models for the outcome model for the NIHSS (Y) and for the nonresponse model (R). For the nonresponse model, in addition to NIHSS, we included all 13 fully observed variables to fit a logistic regression model to predict probability of missingness. For the outcome model, we used the same covariates in the nonresponse model to fit a linear regression model. To compute the NMI and NBI estimators with different weighting schemes, we chose $K = 3$ nearest neighbor and ω_1 and ω_2 varied as 0.2,0.5, and 0.8 as in the simulation studies. In NMI we used $M = 30$ imputations and used Rubin's combining rules to get estimates for the mean NIHSS. Similarly, we used NBI with $M = 50$ bootstrap imputations and used rules for combining bootstrap imputation (Efron, 1994). Additionally, we also presented the results of the CC, MI assuming MAR, RI, and PPMHD for comparison. For MI, Bayesian linear regression was used with all 12 covariates as predictors of NIHSS. For RI, all 12 complete covariates were used in the imputation and nonresponse models. For PPMHD, all 12 variables were used to create a proxy by regressing on NIHSS to create predicted values for all subjects. We applied the PPMHD with the assumption of MAR ($\lambda = 0$), weak MNAR ($\lambda = 1$), strong MNAR ($\lambda = \infty$), and chose the closeness parameter $c = 3$ as recommended by Sullivan and Andridge (2015).

Table 3.5 provides estimates of the mean NIHSS of the sample for CC, PMI, RI, NBI, NMI, and PPMHD along with standard errors (SE), and 95% confidence intervals (CI). All methods except the PPMHD assuming a strong MNAR assumption ($\lambda = \infty$) produce similar point estimate of the mean NIHSS level. The RI method produced a much higher estimate of SE and a slightly higher estimate of the mean compared to the other methods which was shown in the simulations. The MI, RI, NBI, NMI, and PPMHD methods produce slightly higher SEs compared to the CC analysis. As expected, the mean estimate of NIHSS for MI is similar to the estimate using PPMHD with $\lambda = 1$, since both assume a MAR mechanism. The mean estimates assuming MAR are very close to the estimates obtained from NBI(3, 0.5, 0.5) and NMI(3, 0.5, 0.5). Thus, it is likely that the missing data mechanism is close to MAR in this dataset. Larger weights for the predictive scores of the missing probability and outcome resulted in slightly lower estimates for the mean as compared to the other weighting schemes. The PPMHD$(3, \infty)$ which assumes an extreme MNAR mechanism resulted in much lower mean estimates of NIHSS. However, these differences are not clinically meaningful since NIHSS can range from 0 to 42. One disadvantage of the PPMHD compared to the NBI and NMI is the reliance on one set of predictive means from the mixture models to create imputations. The correlation between the outcome and predicted means among observed values was $\rho = 0.045$, which was not strong making it difficult to find close donors for the unobserved values and explains the lower value. In contrast, the NBI and NMI methods iterate between the nonresponse model and outcome model rather than relying solely on one set of predictive means. Our results seem to indicate the models for Y and for the missing probability are approximately correct. As a result, either NBI(3, 0.5, 0.5) or NBI(3, 0.5, 0.5) estimates could be chosen as the estimate of the mean NIHSS in our sample.

3.5 Discussion

A challenging problem in the analysis of missing data concerns how to handle missing that is missing not at random which results in a non-ignorable missing data mechanism. In the analysis phase of research, one should never completely rule out the possibility of non-ignorability, under which modeling of the missing data mechanism is essential to provide valid inferences. We have developed a nonparametric approach that is useful when the missing data is non-ignorable. Compared to PMI, RI, and PPMHD, our approach is more robust because our proposed weighting scheme allows for a sensitivity analysis so investigators can incorporate prior beliefs on the nonresponse and outcome models. Our approach also incorporates a bootstrap step, which aids in integrating appropriate variability across the imputations. More importantly, we can evaluate the validity of the working models, which in turn enables investigators to choose an optimal imputation estimator. Furthermore, our proposed approach can easily be extended to other settings such as regression analysis with missing data. From simulations studies, we suggest that investigators perform the sensitivity analysis and set ω_2 to a larger value when lacking prior knowledge about the strength of the models. It is of future interest to extend our methods to studies with general missing data patterns.

Table 3.1: Estimating the mean in the presence of non-ignorable missing data with both the nonresponse and outcome model correctly specified

Correlation	Method	RB(%)	SD	SE	MSE	CR(%)
$\rho = 0.5$	CC	4.79	0.183	0.154	0.263	27
	PMI	4.16	0.213	0.215	0.219	53
	RI	9.53	0.405	0.816	1.071	92
	PPMHD(3,1)	3.73	0.252	0.304	0.202	82
	PPMHD(3,∞)	4.19	0.219	0.303	0.224	74
	NBI(3,0.2, 0.8)	1.43	0.305	0.322	0.140	94
	NBI(3,0.5,0.5)	1.67	0.264	0.293	0.127	93
	NBI(3,0.8,0.2)	1.98	0.235	0.268	0.123	92
	NMI(3,0.2, 0.8)	1.49	0.311	0.308	0.118	90
	NMI(3,0.5,0.5)	1.65	0.267	0.278	0.098	89
	NMI(3,0.8,0.2)	1.99	0.237	0.252	0.096	90
$\rho = 0.0$	CC	5.81	0.181	0.153	0.370	14
	PMI	4.26	0.218	0.222	0.229	54
	RI	10.84	0.454	0.920	1.380	91
	PPMHD(3,1)	3.82	0.234	0.307	0.200	80
	PPMHD(3,∞)	5.03	0.199	0.266	0.292	51
	NBI(3,0.2, 0.8)	1.81	0.327	0.364	0.140	95
	NBI(3,0.5,0.5)	2.16	0.284	0.319	0.127	93
	NBI(3,0.8,0.2)	2.48	0.247	0.291	0.123	90
	NMI(3,0.2, 0.8)	1.87	0.332	0.347	0.145	90
	NMI(3,0.5,0.5)	2.14	0.283	0.308	0.126	89
	NMI (3,0.8,0.2)	2.49	0.248	0.274	0.124	87
$\rho = -0.5$	CC	6.35	0.180	0.152	0.435	8
	PMI	4.51	0.223	0.220	0.253	50
	RI	11.44	0.509	0.951	1.567	91
	PPMHD(3,1)	4.45	0.218	0.289	0.246	70
	PPMHD(3,∞)	5.77	0.189	0.231	0.368	30
	NBI(3,0.2, 0.8)	2.25	0.319	0.367	0.152	94
	NBI(3,0.5,0.5)	2.57	0.273	0.326	0.140	93
	NBI(3,0.8,0.2)	2.85	0.242	0.291	0.140	88
	NMI(3,0.2, 0.8)	2.28	0.322	0.354	0.156	90
	NMI(3,0.5,0.5)	2.55	0.278	0.310	0.142	87
	NMI(3,0.8,0.2)	2.93	0.244	0.275	0.145	84

Table 3.2: Estimating the mean in the presence of non-ignorable missing data with both the nonresponse and outcome model incorrectly specified

Correlation	Method	RB(%)	SD	SE	MSE	CR(%)
$\rho = 0.5$ (32%)	CC	-15.255	0.221	0.14	1.538	0
	PMI	3.386	0.219	0.175	0.121	69
	RI	3.136	0.267	0.576	0.133	99
	PPMHD(3,1)	3.507	0.263	0.233	0.147	82
	PPMHD(3,∞)	3.621	0.246	0.225	0.144	75
	NBI(3,0.2, 0.8)	-1.919	0.277	0.272	0.099	94
	NBI(3,0.5,0.5)	-1.542	0.230	0.224	0.067	94
	NBI(3,0.8,0.2)	-1.213	0.215	0.213	0.055	90
	NMI(3,0.2, 0.8)	3.066	0.238	0.201	0.116	80
	NMI(3,0.5,0.5)	4.446	0.264	0.212	0.195	61
	NMI(3,0.8,0.2)	3.293	0.276	0.255	0.145	86
$\rho = 0.0$	CC	2.669	0.128	0.082	0.062	65
	PMI	5.02	0.132	0.121	0.178	12
	RI	4.462	0.168	0.577	0.155	100
	PPMHD(3,1)	6.092	0.134	0.132	0.255	6
	PPMHD(3,∞)	7.688	0.162	0.162	0.404	3
	NBI(3,0.2, 0.8)	0.99	0.263	0.282	0.079	96
	NBI(3,0.5,0.5)	1.30	0.245	0.268	0.077	95
	NBI(3,0.8,0.2)	1.70	0.231	0.259	0.082	94
	NMI(3,0.2, 0.8)	4.954	0.149	0.164	0.179	32
	NMI(3,0.5,0.5)	4.931	0.146	0.16	0.177	3
	NMI(3,0.8,0.2)	4.958	0.146	0.158	0.179	3
$\rho = -0.5$	CC	2.53	0.129	0.082	0.057	67
	PMI	4.92	0.124	0.121	0.170	11
	RI	4.37	0.163	0.571	0.149	1
	PPMHD(3,1)	5.98	0.131	0.132	0.246	6
	PPMHD(3,∞)	7.53	0.159	0.162	0.388	4
	NBI(3,0.2, 0.8)	4.88	0.142	0.187	0.173	42
	NBI(3,0.5,0.5)	4.86	0.139	0.181	0.171	40
	NBI(3,0.8,0.2)	4.86	0.137	0.179	0.170	38
	NMI(3,0.2, 0.8)	4.90	0.142	0.172	0.174	38
	NMI(3,0.5,0.5)	4.91	0.138	0.164	0.173	35
	NMI(3,0.8,0.2)	4.93	0.137	0.16	0.174	32

Table 3.3: Estimating the mean in the presence of non-ignorable missing data with the response model misspecifed

Correlation	Method	RB(%)	SD	SE	MSE	CR(%)
$\rho = 0.5$	CC	-15.59	0.22	0.14	1.50	0
	PMI	-1.265	0.177	0.168	0.042	90
	RI	-1.31	0.237	0.478	0.067	100
	PPMHD(3,1)	-1.53	0.181	0.176	0.048	89
	PPMHD(3,∞)	-0.74	0.186	0.179	0.038	94
	NBI(3,0.2, 0.8)	-6.92	0.374	0.445	0.446	81
	NBI(3,0.5,0.5)	-5.67	0.308	0.382	0.300	83
	NBI(3,0.8,0.2)	-4.67	0.271	0.347	0.213	87
	NMI(3,0.2,0.8)	-2.11	0.269	0.268	0.101	89
	NMI(3,0.8,0.2)	-1.85	0.229	0.224	0.074	91
	NMI(3,0.8, 0.2)	-1.61	0.208	0.209	0.060	91
$\rho = 0.0$	CC	-4.14	0.174	0.110	0.140	50
	PMI	-0.42	0.136	0.135	0.020	94
	RI	-0.69	0.173	0.464	0.033	100
	PPMHD(3,1)	-0.22	0.136	0.136	0.019	95
	PPMHD(3,∞)	0.22	0.139	0.140	0.020	95
	NBI(3,0.2, 0.8)	-0.63	0.164	0.206	0.030	98
	NBI(3,0.5,0.5)	-0.55	0.151	0.186	0.025	98
	NBI(3,0.8,0.2)	-0.54	0.145	0.179	0.023	98
	NMI(3,0.2, 0.8)	-0.61	0.168	0.190	0.031	95
	NMI(3,0.5,0.5)	-0.55	0.152	0.171	0.025	96
	NMI(3,0.8,0.2)	-0.52	0.147	0.165	0.023	96
$\rho = -0.5$	CC	2.53	0.129	0.082	0.057	67
	PMI	-0.25	0.118	0.115	0.014	95
	RI	0.10	0.158	0.465	0.025	100
	PPMHD(3,1)	-0.71	0.119	0.121	0.017	95
	PPMHD(3,∞)	-1.43	0.128	0.130	0.029	88
	NBI(3,0.2, 0.8)	-0.38	0.136	0.182	0.019	99
	NBI(3,0.5,0.5)	-0.23	0.136	0.173	0.019	98
	NBI(3,0.8,0.2)	-0.24	0.127	0.162	0.016	98
	NMI(3,0.2, 0.8)	-0.23	0.150	0.167	0.023	97
	NMI(3,0.5,0.5)	-0.27	0.134	0.155	0.018	96
	NMI(3,0.8,0.2)	-0.27	0.131	0.152	0.018	97

Table 3.4: Estimating the mean in the presence of non-ignorable missing data with the outcome model misspecifed

Correlation	Method	RB(%)	SD	SE	MSE	CR(%)
$\rho = 0.5$	CC	-19.35	0.200	0.135	2.436	0
	PMI	-3.11	0.209	0.205	0.106	78
	RI	NA				
	PPMHD(3,1)	-3.02	0.211	0.215	0.103	79
	PPMHD(3,∞)	-1.51	0.223	0.228	0.064	93
	NBI(3,0.2, 0.8)	-6.92	0.374	0.445	0.445	81
	NBI(3,0.5,0.5)	-5.668	0.308	0.382	0.300	83
	NBI(3,0.8,0.2)	-4.67	0.271	0.347	0.213	87
	NMI(3,0.2,0.8)	-6.94	0.373	0.419	0.447	71
	NMI(3,0.8,0.2)	-5.69	0.309	0.354	0.302	74
	NMI(3,0.8, 0.2)	-4.62	0.276	0.318	0.213	80
$\rho = 0.0$	CC	-7.33	0.152	0.107	0.366	2
	PMI	-1.82	0.201	0.206	0.061	90
	RI	-4.98	0.500	0.808	0.405	99
	PPMHD(3,1)	-0.69	0.218	0.227	0.050	96
	PPMHD(3,∞)	0.84	0.249	0.262	0.066	96
	NBI(3,0.2, 0.8)	-4.22	0.338	0.417	0.228	92
	NBI(3,0.5,0.5)	-3.70	0.302	0.374	0.178	93
	NBI(3,0.8,0.2)	-3.16	0.266	0.343	0.135	95
	NMI(3,0.2, 0.8)	-4.12	0.359	0.379	0.236	83
	NMI(3,0.5,0.5)	-3.69	0.290	0.339	0.170	88
	NMI(3,0.8,0.2)	-3.29	0.269	0.31	0.141	86
$\rho = -0.5$	CC	-0.03	0.131	0.083	0.017	0.94
	PMI	-1.49	0.191	0.190	0.051	91
	RI	NA				
	PPMHD(3,1)	-2.09	0.234	0.243	0.083	89
	PPMHD(3,∞)	-2.72	0.299	0.325	0.137	88
	NBI(3,0.2, 0.8)	-2.07	0.313	0.379	0.126	96
	NBI(3,0.5,0.5)	-2.16	0.281	0.347	0.109	95
	NBI(3,0.8,0.2)	-2.19	0.251	0.318	0.094	95
	NMI(3,0.2, 0.8)	-2.07	0.320	0.356	0.130	94
	NMI(3,0.5,0.5)	-2.19	0.289	0.325	0.114	93
	NMI(3,0.8,0.2)	-2.15	0.262	0.297	0.098	94

Table 3.5: Estimating the mean National Institute of Health Stroke Score

Method	Estimate	SE	95% CI
CC	7.377	0.187	(7.010, 7.745)
PMI	7.124	0.187	(6.743, 7.505)
RI	7.336	0.913	(5.469, 9.203)
PPMHD(3,1)	6.643	0.193	(6.264, 7.022)
PPMHD(3,∞)	6.878	0.241	(6.405, 7.351)
NBI(3,0.2,0.8)	7.008	0.217	(6.582, 7.434)
NBI(3,0.5,0.5)	7.183	0.229	(6.425, 7.759)
NBI(3,0.8,0.2)	7.002	0.229	(6.553, 7.450)
NMI(3,0.2,0.8)	7.067	0.295	(6.464, 7.669)
NMI(3,0.5,0.5)	7.084	0.201	(6.676, 7.493)
NMI(3,0.8,0.2)	7.018	0.185	(6.642, 7.394)

Chapter 4

Evaluating Posterior Predictive Checking For Imputation Models Under the Missing Not at Random Assumption

Abstract

Multiple imputation is a popular method for handling missing data. In multiple imputation, missing values are replaced with several "plausible values" using posterior predictive draws from imputation models. It is important to create a general imputation model that is close to the true model because the validity of the completed data inferences depends on the adequacy of the imputation model. Posterior predictive checking (PPC) has been recommended as a potential method for checking imputation models under the assumption of missing at random by simulating "replicated" data from the posterior predictive distribution of the model. Although PPC has been proposed in the literature for missing at random data, no studies (to the best of our knowledge) have formally evaluated whether PPC is useful for identifying problems with imputation models under the assumption of missing not at random. The aim of our study is to evaluate the performance of PPC as an imputation diagnostic under the assumption of missing not at random. Using simulation studies, we examine whether PPC can reliably identify imputation models that could potentially lead to biased substantive inference. More specifically, we use PPC p-values as our summary diagnostic measure, where extreme p-values (i.e. p-values close to 0 or 1) suggest a misfit between the imputation model and the data. In addition, we use graphical checks for exploring the behavior of the PPC p-values across simulations. We generate a gold standard imputation model and deliberately misspecify imputation models to determine whether PPC is effective in identifying a departure from the true model. The method is illustrated using a real data example from the Georgia Coverdell Acute Stroke Registry.

4.1 Introduction

Multiple imputation (MI) is a popular method for handling missing data which can be partly attributed to its ease of use (Rubin, 1987). MI replaces missing values M times with "plausible values" using posterior predictive draws from imputation models to create M completed data sets. Standard completed-data analyses can be used on each data set completed by imputation, and Rubin's combining rules (Rubin, 1987) are used to create a single inferential summary. In the last few decades, the general framework and statistical theory for MI have been well developed. However, checking imputation models is not common in practice, although it is generally an important part of any statistical procedure. In addition, most imputation diagnostic methods assume that the data are missing at random [MAR] (Rubin, 1976) which implies that given the observed data, data are missing independently of unobserved data. A less restrictive assumption is that the missing observations are related to values of unobserved data referred to in the literature as missing not at random (MNAR). To the best of our knowledge, there are no methods developed to handle MNAR imputation diagnostics.

Meng (1994a) suggested that the imputation model be congenial or general enough to preserve any associations among variables that may be the target in the subsequent, completed data analyses. Furthermore, a general imputation model that is close to the true model allows for accommodation of a wide range of statistical models that can be used on the completed data sets. In order to construct a reasonable, general imputation model, a major issue is to not exclude any important predictors or relationships (i.e. nonlinear relationships) among the data. Excluding important features may lead to imputation models that are not as general than the subsequent analysis and can potentially bias the results. Diagnostic and model checking of imputation models is a natural way to determine whether these assumptions hold.

Although diagnostic methods are scarce, diagnostic testing for missing data models have a long history. One of the first studies for diagnostic for missing values was introduced by Poirier and Ruud (1983). They introduced a more general case of the Heckman (1977) selection model to handle missing data in a maximum likelihood based approach. Violation of either homoscedasticity and lognormality could potentially result in inconsistency of the estimators in those models. To identify departures from model assumptions, they proposed the use of Lagrange multipliers test. However, they did not study diagnostic for imputations because it was before (Rubin, 1987) published his pioneering work on MI to handle nonresponse in surveys.

Previous work on imputation diagnostic is limited to the assumption that the data are missing at random. For instance, Raghunathan and Bondarenko (2007) proposed the use of propensity scores as a diagnostic tool to check the validity of imputed vales in MI. They checked the equality of the distributions of the observed and missing values conditional on the response propensity score. Wang (2010) extended this diagnostic approach to include a regression of both the observed and imputed data as a function of the predicted propensity score and the missingness indicator. With the extension, Wang (2010) was able to check whether the imputation model used to generate imputations would preserve the associations among variables in the dataset by determining if the missingness was completely explained by the response propensity score.

Several authors have used graphical tools and numerical test such as Kolmogorov-Smirnov (KS) test to assess plausibility of imputations. For example, Abayomi et al. (2008) examined the empirical density plots, bivariate scatter plots, and residual plots to identify dramatic differences from the observed and imputed data. Bondarenko and Raghunathan (2016) made graphical comparisons of the observed and imputed values conditional on the response propensity to assess the suitability of imputations from imputation models. Abayomi et al. (2008) and Nguyen et al. (2013) used the KS test to diagnose problems with imputation models by comparing the empirical distribution of the observed and imputed data. Abayomi et al. (2008) flagged imputed variables with statistically significant differences and further examined variables using graphical techniques. Nguyen et al. (2013) suggested that the imputed variables required more rigorous evaluation after the KS test is performed. Nguyen examined the behavior of the KS p-value under various scenarios in simulations, including varying the amount of missing data, misspecified imputation models, and skewed and heavy-tailed distributions.

He and Zaslavsky (2012) and Nguyen et al. (2015) used a diagnostic method based on posterior predictive checking [PPC] (Gelman et al., 1996), namely the posterior predictive p-value (Meng, 1994b), to determine the adequacy of imputation models by applying subsequent analyses of interest to both the completed data and their posterior replicates simulated under the imputation model. Large differences between the estimates using the completed data and the simulated replicates may suggest model inadequacy. He and Zaslavsky (2012) and Nguyen et al. (2015) checked imputation models assuming the missing data are MAR. However, principled diagnostic approaches that can handle the less restrictive assumption of MNAR have not been investigated in the literature. Motivated by such facts, our primary goal is to evaluate whether the diagnostic methods of He and Zaslavsky (2012) and Nguyen et al. (2015) are applicable in the case of MNAR imputation models.

The remainder of this paper is organized as follows. In Section 2, we review several approaches for multiple imputation developed assuming the data are MNAR. In Section 3, we review the posterior predictive checking methods of Gelman et al. (1996) and demonstrate its application to MNAR imputation models. In Section 4, we perform simulations to evaluate the performance of PPC p-values as a summary diagnostic measure for MNAR imputation models. In Section 5, we apply the approach to imputation models using the Georgia Coverdell Acute Stroke registry (GCASR) data. In Section 6, we conclude with a brief discussion of future research directions.

4.2 Missing Not at Random Imputation Methods

Suppose data are collected for n samples but some variables in the study are subjected to missingness. We let Y denote the variable with missing values, R denote the response indicator which takes values 1 if Y is observed and 0 if Y is missing, and $\boldsymbol{X}$ denote a set of fully observed auxillary variables that are predictive of either Y or R. For a given sample we can partition Y into two separate components; the observed observations of Y are denoted by Y_{obs} and the missing observations are denoted by Y_{mis}. Further suppose the probability that Y is missing depends upon the missing value itself, thus the data are MNAR. The development of methods for the analysis of data under the assumption of MNAR has been an active area of research. In this paper, methods that incorporate imputation models are of particular interest, since PPC is only applicable to imputation models. Methods that involve imputation models to impute MNAR data include the random indicator method (RI, (Jolani, 2012)) and the proxy-pattern mixture hot deck (PPMHD, (Sullivan and Andridge, 2015)). In this section, we briefly review the aforementioned imputation methods for handling data that are MNAR.

4.2.1 Random Indicator Method

Jolani (2012) proposed drawing a pseudo response indicator from the model for the missing data mechanism using the selection modeling approach. Under the selection modeling approach, the joint distribution of R and Y is specified through models for the marginal distribution of Y and the conditional distribution of R given Y. Jolani (2012) iteratively drew imputations for the incomplete variable Y and the realization of the response indicator R, denoted R^*. Initial values for Y_{mis} are randomly drawn from Y_{obs}. Then, a parametric model for the missing data mechanism is fitted, $E(R|Y,X) = f_1(X,Y;\alpha)$, where f_1 is a real-valued smooth function, such as the logit, and α^* can be drawn from its posterior distribution given the estimated value $\hat{\alpha}$. Given $\hat{\alpha}$, the pseudo random indicator can be drawn from a Bernoulli process where the probability is $P(R=1|X,Y^{(t)})$, $Y^{(t)} = (Y_{obs}, Y_{mis}^{imp,m^{(t)}})$, and t denotes an iteration of the algorithm, where M is the number of imputations. In addition, using the observed data, a model for the outcome Y, $E(Y|X,R=1,R^*=r^*) = X\beta + \delta(r^*-1)$, is fitted where r^* denotes the value of the pseudo indicator which takes values 0 or 1, β is a vector of regression coefficients, and δ is the adjustment parameter. By cross-classifying Y by R and R^*, various properties on the distribution of the missing data are obtained through the adjustment parameter. Given the prior for β, β^* is drawn from its posterior distribution. An update for the missing values are created conditional on the pseudo random indicator R^*, such that when $r=1$, we predict Y_{mis} using $X\beta^* - \hat{\delta}$, and using $X\beta^* - 2\hat{\delta}$ when $r=0$. The algorithm is repeated until convergence and the last values are treated as one imputation for the missing values. M imputations for Y_{mis} are generated by starting with new initial values for Y_{mis}.

4.2.2 Proxy Pattern Mixture Hot deck

Sullivan and Andridge (2015) proposed an approach to impute MNAR data through the use of the hot deck. Predicted values for Y are based on a pattern-mixture model (Little, 1993), where the joint distribution of Y and R are specified as the marginal distribution of R and the conditional distribution of Y given R. A parametric model, regressing Y on fully observed covariates X, is used to create predicted means for both observed values and missing values under varying assumptions of the missing data mechanism. The sensitvity parameter λ determines the missingness mechanism; the values λ are varied as 0 to assume a MAR mechanism, 1 for a weak MNAR mechanism, and $\lambda = \infty$ indicating an extreme case of MNAR. For a given assumption on the missingness mechanism, the predicted means are used to define distances between observed values and unobserved values and probabilities of selection proportional to those distances. An imputation is created by randomly selecting an observed value for the missing values using the selection probabilities.

4.3 Diagnostic Methods

4.3.1 Posterior Predictive Checking

Our goal is to evaluate whether posterior predictive checking will uncover discrepancies in statistical inferences when aspects of the MNAR imputation model do not fit the data. Posterior predictive checking compares the observed Y with draws of replicates denoted Y^{rep} that are drawn from their posterior predictive distribution of $Y = (Y_{obs}, Y_{mis})$ using a discrepancy function denoted by Q. The Q is a scalar function of the data, for example, the mean or median. The posterior predictive p-value(Meng, 1994b) is a commonly used diagnostic measure in posterior predictive checking. The posterior predictive p-value is the probability that the replicated data would be more extreme than the observed data. Meng (1994b) define the posterior predictive p-vlaue for an imputation model as

$$\begin{aligned} p_{B,com} &= P(Q(Y^{rep}) \geq Q(Y_{obs}, Y_{mis})|Y_{obs}, X) && (4.1) \\ &= \int\int I(Q(Y^{rep}) \geq Q(Y_{obs}, Y_{mis}) f(Y^{rep}, Y_{mis}|Y_{obs}, X) dY^{rep} dY_{mis} && (4.2) \end{aligned}$$

where $I(.)$ is the indicator function and the replicated data is $Y^{rep} = (Y^{rep}_{obs}, Y^{rep}_{mis})$. For $d = 1, ..., D$,we use the existing imputation methods for MNAR data introduced in Section 4.2, namely RI and PPMHD, to impute $Y^{imp,d}_{mis}$ and simulate the replicated data. We can estimate p_{post} by simulation as the proportion of D draws for which $Q(Y^{rep,d}) \geq Q(Y_{obs}, Y_{mis}^{imp,d})$. Hence comparing the realized test quantities $Q(Y_{obs}, Y_{mis}^{imp,d})$ with its simulated replicates $Q(Y^{rep,d})$. The goal of posterior predictive checking is to assess whether components of the data that do not fit the imputation model will lead to discrepancies. He and Zaslavsky (2012) showed that extreme p values, either close to 0 or 1, would suggest there are discrepancies between (Y_{obs}, Y_{mis}^{imp}) and $Y^{rep} = (Y_{obs}^{rep}, Y_{mis}^{rep})$ that may not be fully explained by chance under MAR. Our methods are distinctive from He and Zaslavsky (2012), in that our primary objective is to determine whether this diagnostic approach for imputation models will detect discrepancies assuming the missing data mechanism is non-ignorable. Under nonignorable missingness, one assumes the response mechanism depends on unobserved values, and the outcome model for the variable with missing values and the response model must be modeled jointly, thus two models are used to conduct imputation which differentiates our approach from the previous diagnostic approach of He and Zaslavsky (2012). If PPC is applicable in this setting, then we expect the posterior predictive p-value of a correctly specified imputation model to be closer to 0.5 than a misspecified imputation model. Hence, an incorrect imputation model can be flagged for further review of the imputation model.

Variance is added to the comparison of Y_mis and Y_{mis}^{rep} since it is generated from the same imputation model but imputed separately. Adding variance to the comparison reduces the power. He and Zaslavsky (2012) reduce the variance of the distribution of the completed data discrepancy and use the posterior predictive p-value corresponding to this expected completed-data discrepancy, denoted by $p_{B,ecom}$ where

$$p_{B,ecom} = P(E\left[Q(Y_{com}^{rep})|Y_{obs}^{rep}, Y_{obs}\right] \geq E\left[Q(Y_{obs}, Y_{mis})|Y_{obs}^{rep}, Y_{obs}\right]|Y_{obs}).$$

4.3.2 Discrepancy functions targeted to substantive inferences

The aim in the statistical analysis phase of research is to obtain the substantive analysis results or the analyses that would be performed in the absence of missing data. Yet, missing data must be addressed before the substantive analysis can be performed. He and Zaslavsky (2012) suggested to use discrepancy functions Q that are estimands of the target analyses in posterior predictive checking instead of using generic summary measures. Using these targeted estimands connects the model diagnostics with the analytic objectives. For example, if the scientific objective of the analysis is a simple linear regression, then a discrepancy function that is an estimand of the analysis might include the regression coefficients. In simulation, He and Zaslavsky (2012) use the mean, variance, and regression estimates as discrepancy functions. We were also interested in examining other estimands such as the t-statistics. They argue that model diagnostics targeted to statistical inferences may be more appropriate to detect relevant imputation model deficiencies. We also consider it may be more appropriate to apply target analyses of interest to both the completed data and their posterior replicates simulated under the imputation model to detect model discrepancies in the case of MNAR data.

4.4 Simulations

4.4.1 Goal

We conducted a simulation study in order to assess the performance of posterior predictive checking to diagnose the adequacy of imputation models in the presence of nonignorable missingness. In addition to examining the p-values, we also sought to compare its performance across different imputation methods, proportion of missingness, and different choices of imputation models and target analyses. We apply RI (Jolani, 2012) and PPMHD (Sullivan and Andridge, 2015) to construct imputation models and impute the missing data and simulate replicates. The posterior predictive p-values (Meng, 1994b) and posterior predictive p-value corresponding to the expected completed-data discrepancy (He and Zaslavsky, 2012) is used as a summary measure to assess the plausibility of the posited imputation models. We deliberately misspecify the imputation model to determine whether PPC are effective in identifying imputation model inadequacies. If the imputation model is correctly specified and the diagnostic measure is appropriate assuming the data are nonignorably missing, then we expect the posterior predictive p-value to be closer to 0.5 than any alternative model. Alternatively, if the imputation model is poorly specified, such as through the omission of important variables, this can lead to biased results and we expect to have extreme p-values (i.e. close to 0 or 1). Extreme p-values suggest deficiencies in the imputation model.

4.4.2 Simulation setting

We adopt the simulation set-up of He and Zaslavsky (2012) but to include imputaitons for nonignorable missingness and additional discrepancy functions. In our simulations, we also consider two variables, a fully observed Y_1 and incomplete Y_2 with $n = 1000$. The data generating model is $Y_1 \sim Unif(-3, 3)$, $Y_2|Y_1 \sim N(1 + Y_1 + 0.5Y_1^2, 1)$. We use 300 Monte Carlo datasets in simulation. The following describes the simulation design:

- *True Response model*: the response model depends on Y_2 therefore is missing not at random with $logit[P(Y_2 \text{ is missing}|Y_1, Y_2)] = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2$ where $\boldsymbol{\beta} = (-3.1, 0.5, 0.4)$ to induce 20% missingess and where $\boldsymbol{\beta} = (1.3, 0.5, 0.4)$ induces 80% of missingness on Y_2. The probability of response depends on the variable with missing values, hence the missingness mechanism is nonignorable.

- *Working model for the outcome and response*:

  a. Correct imputation model: Y_2 is specified as a quadratic function of Y_1

  b. Incorrect imputation model: Y_2 is specified as a linear function of Y_1

In the latter, we deliberately misspecify the aforementioned response and outcome imputation models to determine whether the diagnostic approach will be able to identify discrepancies in imputation models in the presence of nonignorable missingness.

We use discrepancy functions from completed-data analyses and specify the following discrepany functions Q.

I. Mean of the incomplete variable Y_2

II. Variance of the incomplete variable Y_2

III. The coefficient estimates and t-statistics for a linear and quadratic regression of Y_2 and Y_1.

IV. The coefficient estimates and t-statistics for a linear and quadratic regression of Y_1 and Y_2.

We applied the PPMHD using $? = (1, \infty)$ to impute the missing values implying assuming a weak and strong MNAR missing data mechanism for $\lambda = 1$ and $\lambda = \infty$. In addition, we impute missing values using the RI method Jolani (2012). We estimate the completed-data discrepancies p_B, com by simulation with $D_1 = 500$. We also calculate $p_B, ecom$ with $D_2 = 500$ and $D_2 = 20$ and the mean and median posterior predictive p-value of the completed-data and expected completed-data across simulations. Furthermore, we calculate the proportion of times the posterior predictive p-values are closer to 0.5 for the correctly specified imputation model.

4.4.3 Simulation Results

Table 4.1, 4.2, 4.7, and 4.8 show the results using the completed-data discrepancy when data are imputed using RI. Table 4.1 and 4.2 show the results with 20% missingness rate where Table 4.1 displays the mean of the posterior predictive p-values and Table 4.2 displays the median. Table 4.7 and 4.8 show the results with 80% missingness rate where Table 4.7 displays the mean of the posterior predictive p-values and Table 4.8 displays the median. Likewise, Table 4.3, 4.4, 4.9, and 4.10 show the results using the completed-data discrepancy when data are imputed using the PPMHD with $\lambda = 1$ invoking a weak MNAR imputation. Table 4.3 and 4.4 show the results with 20% missingness rate where Table 4.3 displays the mean of the posterior predictive p-values and Table 4.4 displays the median. Table 4.9 and 4.10 show the results with 80% missingness rate where Table 4.9 displays the mean of the posterior predictive p-values and Table 4.10 displays the median. Lastly, Table 4.5, 4.6, 4.11, and 4.12 show the results using the completed-data discrepancy when data are imputed using the PPMHD with $\lambda = \infty$ invoking a strong MNAR imputation. Table 4.5 and 4.6 show the results with 20% missingness rate where Table 4.5 displays the mean of the posterior predictive p-values and Table 4.6 displays the median. Table 4.11 and 4.12 show the results with 80% missingness rate where Table 4.11 displays the mean of the posterior predictive p-values and Table 4.12 displays the median. We present the average of the target analysis statistics using the completed data and their replicates, namely $\bar{Q}(Y_{obs}, Y_{mis})$ and $\bar{Q}(Y_{oom}^{rep})$.

4.4.3.1 Effect of imputation model and analysis models

When the imputation model is quadratic thus matches the data-generating model, the completed-data posterior predictive p-value associated with the quadratic imputation model is generally closer to 0.5. However, there is an exception when the missing data are imputed using the PPMHD with $\lambda = \infty$ and the average percent missing is 80.

Also across simulations and different methods, the proportion of times the posterior predictive p-value is closer is 0.5 than the linear imputation model is 0.79. Hence, PPC may be effective in identifying deficiencies in imputation models.

We use discrepancy functions Q targeted to analytic inferences including the mean, variance, linear and quadratic regressions. Diagnostic results for the regression estimates are similar to those of the t-statistics in terms of resulting posterior predictive p-value. For the quadratic regression of Y_2 on Y_1 assuming a linear imputation model, the coefficients and t-estimates for the intercept and quadratic term are very different between the completed data and their replicates. As a result, it is evident PPC may help to identify the curvature that is omitted from the imputation model. The corresponding $p_{B,com}$ values are also rather extreme suggesting model inadequcy for the linear impuation model.

4.4.3.2 Effect of imputation method and proportion of missingness

When the percent of missing values is small, or about 20% the posterior predictive p-values associated with the linear imputation model are generally more extreme. As a result, there is evidence to suggest the discrepancy is unlikely to be due to chance. Extreme p-values reflect the implausibility of the imputation model and therefore suggest examining other models. However, when the percent of missing values is about 80%, the posterior predictive p-values associated with the linear imputation model are not as extreme. In addition, the average discrepancies for some of the target analyses are substantial when the missing data are imputed using PPMHD assuming $\lambda = \infty$, therefore showing its more difficult to recover the distribution of the incomplete variable Y_2 with higher missing rate in this setting.

4.4.3.3 Effect of mean and median posterior predictive p-values

The mean completed-data posterior predictive p-value and the mean expected completed-data posterior predictive p-value was computed as in He and Zaslavsky (2012). Additionally, we calculated the median of the p-values but did not see a considerate difference between the two summary measures. The results also show using the expected completed-data discrepancy show similar trends to those using the completed-data discrepancy.

4.5 Applications to Stroke Registry Data

We present an example that demonstrates the use of PPC p-values for diagnosing discrepancies in non-ignorable nonresponse imputation models. These data were obtain from the Georgia Coverdell Acute Stroke Registry (GCASR). A detailed description of the registry data set was reported elsewhere (Camp et al., 2015). The primary goal of the GCASR program is to improve the care of acute stroke patients in the hospital and pre-hospital setting. We used a sample of 2,115 patients with acute ischemic stroke in which data were collected from multiple sources, including patient surveys and medical records, between 2007 and 2011. These data include fully observed variables age, gender, race, and dichotomous variables for obesity, diabetes, atrial fibrillation, hypertension, IV tpa, and whether or not the patient had insurance. Other clinical variables collected for all patients were cholesterol level, systolic blood pressure, diastolic blood pressure, and hospital. While there were many variables with missing values, for illustrating the diagnostic method we focus on imputation of the National institute of Health stroke score (NIHSS). The NIHSS is a clinical assessment tool to measure stroke severity, predict patient clinical outcome, and is often used to determine appropriate treatment. Of the 2115 patients, 681 were missing NIHSS (32% missing). Clinicians often believe the assessment is time-consuming and typically do not calculate the score if they do not have reason to believe the patient is a candidate for acute treatment with intravenous thrombolytics or enrollment in investigational protocols (Richardson et al., 2006). Furthermore, these barriers in implementation of the score created concern for whether or not the data were nonignorably missing.

The goal of our application of PPC is to examine the ability to detect imputation model deficiencies assuming non-ignorable missingness. To account for a potential non-ignorable missing data mechanism, we applied RI,PPMHD,NBI,and NMI to estimating the overall mean NIHSS in the study population. Clinicians believe the thirteen fully observed variables are potentially predictive of NIHSS or the probability of missingness, therefore we use all the fully observed variables to impute the incomplete variable NIHSS. In addition, we investigate imputing NIHSS using only systolic blood pressure of the patient as a variable in the imputation model. Meng (1994a) suggested imputation models be general to accommodate a wide range of statistical analyses that may be conducted using multiply imputed data sets.The former imputation model is more general, but the latter is much more restrictive and may not be congenial to completed-data analyses.

We use the mean as the discrepancy function after imputation of NIHSS with the two imputation methods. In Table 4.13, we present the average discrepancy of (Y_{obs}, Y_{mis}) and its simulated replicate (Y_{com}^{rep}). We perform PPC using both completed data $(D = 500)$ and expected completed-data discrepancies $(D = 500$ and $D_2 = 20)$ by simulation and calculate the associated p-values, p_{com} and p_{ecom} respectively. We use a general imputation model (correct model) that includes 13 variables to use for imputation and also a more restrictive imputation model that only includes one variable systolic blood pressure (misspecified model). Table 4.6 shows that the magnitude of the average discrepancies is comparable when the RI imputation method is used to impute the missing values. In addition, we note that both $p_{B,com}$-values and $p_{B,ecom}$-values are close to 0.5 when using the RI method. We examine more extreme p-values closer to 0 or 1 with the incorrectly specified imputation model 1. More extreme p-values suggest a deficiency in the imputation incorrectly specified model 2. We also applied the PPMHD with the assumption of weak MNAR($\lambda = 1$) and strong MNAR ($\lambda = \infty$) (Sullivan and Andridge, 2015). Alternatively, the posterior predictive p-value is near one when the imputations are from a correctly specified imputation models. Both imputation models appear to have some misfit for the data when imputations are made using PPMHD, yet do not have substantial impact on the imputation inference. However, in simulations in Chapter 3, we showed improved performance of NBI and NMI methods as compared the existing methods suggesting it is more appropriate for the stroke data.

4.6 Discussion

Checking imputation models is not common in practice. Some methods have been developed to assess the adequacy of imputation models which include graphical diagnostics, KS test, and PPC. PPC was shown to be preferable to methods that focus on the plausibility of imputations, because it checks the fit of the model with respect to quantities of scientific interest (Nguyen et al., 2015; He and Zaslavsky, 2012) therefore we focus on PPC as a model diagnostic approach in this paper. To the best of our knowledge, there is no studies evaluating diagnostic methods for imputation assuming the data are MNAR. We wanted to determine whether using posterior predictive p-values with discrepancy functions linked to the target analysis was effective in diagnosing the adequacy of MNAR imputation model in supporting specific subsequent analyses. Our extensive simulations suggest that, in the settings we evaluated, posterior predictive p-values can be useful in diagnosing deficiencies in non-ignorable imputation models.

Table 4.1: Simulation results for mean completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the Random Indicator method. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.07	1.90	0.67	0.67	2.41	2.35	0.60	1	0.96
Var of Y_2	3.59	3.66	0.39	0.35	5.35	5.42	0.42	1	0.73
Linear regression Y_1 on Y_2									
Intercept coefficient	-1.11	-1.04	0.33	0.33	-1.28	-1.24	0.40	1	0.98
linear coefficient	0.54	0.55	0.37	0.54	0.53	0.53	0.55	1	0.99
Intercept t-estimate	-17.2	-16.8	0.38	0.33	-22.9	-22.5	0.40	1	0.73
linear t-estimate	23.3	24.4	0.36	0.38	31.8	32.0	0.44	1	0.99
Linear regression Y_2 on Y_1									
Intercept coefficient	2.07	1.90	0.67	0.67	2.41	2.34	0.60	1	0.96
linear coefficient	0.65	1.90	0.67	0.36	0.95	0.96	0.42	1	0.96
Intercept t-estimate	42.9	39.74	0.67	0.67	46.7	45.3	0.60	1	0.97
Linear t-estimate	23.3	24.4	0.36	0.38	31.7	31.9	0.43	1	0.99
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-1.03	-1.04	0.52	0.50	-1.16	-1.13	0.40	1	0.21
Linear coefficient	0.40	0.55	0.00	0.00	0.38	0.39	0.43	1	1
Quadratic coefficient	0.03	0.00	1.00	1.00	0.02	0.02	0.57	1	1
Intercept t-estimate	-14.4	-16.3	0.88	0.90	-18.2	-18.2	0.55	1	1
Linear t-estimate	7.53	14.0	0.00	0.00	8.52	9.00	0.41	1	1
Quadratic t-estimate	2.87	-0.02	1	1	3.70	3.53	0.56	1	1
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.20	1.90	0.06	0.06	0.97	0.89	0.60	1	1
Linear coefficient	0.65	0.67	0.37	0.37	0.95	0.96	0.42	1	0.70
Quadratic coefficient	0.29	0.00	1	1	0.48	0.48	0.44	1	1
Intercept t-estimate	19.3	26.4	0.10	0.18	20.5	19.0	0.60	1	1
Linear t-estimate	27.2	24.3	0.78	0.98	51.9	52.7	0.41	1	1
Quadratic t-estimate	18.8	0.02	1	1	40.8	41.3	0.42	1	1

Table 4.2: Simulation results for median completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the Random Indicator method. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.06	1.91	0.67	0.67	2.41	2.34	0.60	1	0.96
Var of Y_2	3.59	3.66	0.39	0.35	5.34	5.41	0.41	1	0.73
Linear regression Y_1 on Y_2									
Intercept coefficient	-1.11	-1.04	0.33	0.33	-1.28	-1.24	0.40	1	0.98
Linear coefficient	0.54	0.55	0.37	0.54	0.53	0.53	0.56	1	0.99
Intercept t-estimate	-17.2	-16.8	0.38	0.33	-22.9	-22.5	0.40	1	0.73
Linear t-estimate	23.3	24.4	0.36	0.38	31.8	32.0	0.43	1	0.99
Linear regression Y_2 on Y_1									
Intercept coefficient	2.06	1.90	0.67	0.67	2.40	2.34	0.60	1	0.96
Linear coefficient	0.65	0.67	0.36	0.36	0.95	0.96	0.42	1	0.96
Intercept t-estimate	42.9	39.8	0.67	0.67	46.5	45.4	0.60	0.43	0.97
Linear t-estimate	23.3	24.4	0.36	0.38	31.9	32.1	0.43	1	0.99
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-1.03	-1.05	0.52	0.50	-1.16	-1.13	0.40	1	0.21
Linear coefficient	0.40	0.55	0	0	0.38	0.38	0.43	1	1
Quadratic coefficient	0.03	0.00	1	1	0.02	0.02	0.56	1	1
Intercept t-estimate	-14.4	-16.3	0.89	0.91	-18.2	-18.2	0.55	1	1
Linear t-estimate	7.49	14.0	0.00	0.00	8.45	8.99	0.42	1	1
Quadratic t-estimate	2.84	-0.05	1	1	3.70	3.54	0.55	1	1
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.20	1.90	0.06	0.06	0.97	0.90	0.60	1	1
Linear coefficient	0.65	0.67	0.36	0.36	0.95	0.96	0.42	1	0.70
Quadratic coefficient	0.29	0.00	1	1	0.48	0.48	0.43	1	1
Intercept t-estimate	19.2	26.4	0.10	0.18	20.6	19.0	0.60	1	1
Linear t-estimate	27.2	24.4	0.79	0.98	51.8	52.6	0.41	1	1
Quadratic t-estimate	18.7	0.02	1	1	40.6	41.2	0.42	1	1

Table 4.3: Simulation results for mean completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = 1$ implying a weak MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.36	2.76	0.00	0.00	2.43	2.48	0.11	0.09	1
Var of Y_2	5.31	6.28	0.00	0.00	5.50	5.69	0.17	0.11	1
Linear regression Y_1 on Y_2									
Intercept coefficient	-1.11	-1.26	0.99	0.99	-1.25	-1.26	0.55	0.57	1
Linear coefficient	0.47	0.46	0.73	0.60	0.52	0.51	0.77	0.79	0.40
Intercept t-estimate	-18.2	-20.8	0.97	0.99	-22.0	-22.0	0.50	0.51	1
Linear t-estimate	25.3	28.0	0.08	0.03	30.6	30.5	0.53	0.52	1
Linear regression Y_2 on Y_1									
Intercept coefficient	2.36	2.77	0.00	0.00	8.72	8.78	0.21	0.18	1
Linear coefficient	0.83	0.95	0.00	0.00	2.32	2.19	0.41	0.42	1
Intercept t-estimate	41.6	46.8	0.00	0.00	45.6	45.8	0.44	0.42	1
Linear t-estimate	25.36	28.0	0.08	0.03	30.7	30.6	0.53	0.52	1
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-1.04	-1.23	0.99	1.00	-1.13	-1.12	0.43	0.41	1
Linear coefficient	0.37	0.41	0.28	0.31	0.36	0.33	0.73	0.76	0.44
Quadratic coefficient	0.01	0.01	0.84	0.78	0.02	0.03	0.33	0.29	0.87
Intercept t-estimate	-14.8	-17.1	0.94	0.97	-17.44	-17.3	0.44	0.43	1
Linear t-estimate	7.53	8.58	0.23	0.25	8.03	7.46	0.70	0.72	0.65
Quadratic t-estimate	2.13	1.11	0.81	0.76	3.58	4.05	0.31	0.28	0.80
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.39	2.21	0.00	0.00	1.06	1.10	0.25	0.24	1
Linear coefficient	0.83	0.95	0.01	0.00	0.94	0.95	0.33	0.28	1
Quadratic coefficient	0.32	0.18	1	1	0.46	0.46	0.45	0.43	1
Intercept t-estimate	18.7	26.0	0.00	0.00	19.3	19.6	0.43	0.40	1
Linear t-estimate	29.3	29.5	0.50	0.22	45.2	44.7	0.57	0.58	0.47
Quadratic t-estimate	17.8	9.08	1	1	34.1	33.4	0.60	0.64	1

Table 4.4: Simulation results for median completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = 1$ implying a weak MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.36	2.77	0.00	0.00	2.42	2.47	0.11	0.08	1
Var of Y_2	5.31	6.28	0.00	0.00	5.48	5.68	0.17	0.10	1
Linear regression Y_1 on Y_2									
Intercept	-1.11	-1.26	0.99	1	-1.25	-1.26	0.55	0.57	1
Linear coefficient	0.47	0.46	0.74	0.62	0.52	0.51	0.78	0.79	0.40
Intercept t-estimate	-18.2	-21.0	0.97	0.99	-22.1	-22.1	0.50	0.50	1
Linear t-estimate	25.5	28.1	0.07	0.02	30.7	30.7	0.53	0.53	1
Linear regression Y_2 on Y_1									
Intercept coefficient	2.36	2.77	0.00	0.00	2.45	2.50	0.13	0.10	1
Linear coefficient	0.83	0.96	0.00	0.00	0.96	0.97	0.36	0.32	1
Intercept t-estimate	41.6	46.7	0.00	0.00	45.6	46.0	0.44	0.42	1
Linear t-estimate	24.5	28.1	0.07	0.02	30.7	30.7	0.52	0.52	1
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-1.04	-1.23	0.99	0.99	-1.13	-1.12	0.42	0.40	1
Linear coefficient	0.37	0.41	0.29	0.31	0.36	0.33	0.74	0.76	0.44
Quadratic coefficient	0.01	0.01	0.85	0.79	0.02	0.02	0.32	0.28	0.87
Intercept t-estimate	-14.8	-17.1	0.95	0.98	-17.5	-17.3	0.44	0.43	1
Linear t-estimate	7.50	8.54	0.23	0.23	8.01	7.35	0.71	0.73	0.65
Quadratic t-estimate	2.20	1.22	0.81	0.77	3.56	4.03	0.30	0.26	0.80
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.38	2.20	0.00	0.00	1.06	1.11	0.25	0.24	1
Linear coefficient	0.83	0.96	0.00	0.00	0.94	0.96	0.32	0.27	1
Quadratic coefficient	0.32	0.19	1	1	0.46	0.46	0.44	0.43	1
Intercept t-estimate	18.7	26.1	0.0	0.00	19.3	19.5	0.43	0.40	1
Linear t-estimate	29.3	29.5	0.47	0.19	45.6	45.1	0.57	0.58	0.47
Quadratic t-estimate	17.9	9.20	1	1	34.4	33.7	0.60	0.65	1

Table 4.5: Simulation results for mean completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = \infty$ implying a strong MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.10	2.78	0.00	0.00	2.39	2.50	0.02	0.01	0.96
Var of Y_2	4.35	6.10	0.00	0.00	5.40	5.77	0.05	0.02	1
Linear regression Y_1 on Y_2									
Intercept coefficient	-0.88	-0.83	0.40	0.30	-1.18	-1.18	0.51	0.51	0.75
Linear coefficient	0.42	0.30	1	1	0.49	0.47	0.93	0.96	1
Intercept t-estimate	-13.1	-11.4	0.20	0.14	-20.1	-19.8	0.40	0.38	0.93
Linear t-estimate	18.3	15.2	0.89	0.92	28.0	27.4	0.65	0.67	0.98
Linear regression Y_2 on Y_1									
Intercept coefficient	2.10	2.78	0.00	0.00	2.38	2.49	0.02	0.01	0.01
Linear coefficient	0.60	0.61	0.46	0.45	0.89	0.90	0.32	0.26	0.40
Intercept t-estimate	36.5	39.3	0.06	0.05	43.4	43.6	0.45	0.41	0.99
Linear t-estimate	17.8	14.4	0.91	0.94	28.0	27.4	0.65	0.67	0.96
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-0.80	-0.84	0.64	0.61	-1.07	-1.04	0.33	0.27	0.47
Linear coefficient	0.30	0.31	0.45	0.40	0.35	0.29	0.87	0.91	0.03
Quadratic coefficient	0.02	0.00	0.97	1	0.02	0.03	0.21	0.15	0.99
Intercept t-estimate	-10.6	-9.84	0.34	0.30	-16.0	-15.4	0.30	0.26	0.47
Linear t-estimate	5.44	5.53	0.50	0.30	7.51	6.28	0.85	0.90	0.03
Quadratic t-estimate	2.24	-0.32	0.97	1	3.22	4.09	0.19	0.13	0.98
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.48	3.38	0.00	0.00	1.17	1.30	0.06	0.04	1
Linear coefficient	0.60	0.61	0.46	0.45	0.89	0.91	0.32	0.26	0.43
Quadratic coefficient	0.21	-0.20	1	1	0.40	0.40	0.57	0.59	1
Intercept t-estimate	18.2	33.4	0.00	0.00	18.4	19.0	0.32	0.27	1
Linear t-estimate	19.3	15.6	0.94	0.98	36.7	34.9	0.74	0.79	0.99
Quadratic t-estimate	10.4	-7.92	1	1	25.8	23.8	0.79	0.87	1

Table 4.6: Simulation results for median completed-data discrepancy with $n = 1000$, with 20% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = \infty$ implying a strong MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	2.10	2.88	0.00	0.00	2.38	2.50	0.01	0.00	0.96
Var of Y_2	4.33	6.09	0.00	0.00	5.40	5.78	0.04	0.01	1
Linear regression Y_1 on Y_2									
Intercept coefficient	-0.88	-0.84	0.41	0.29	-1.18	-1.18	0.51	0.50	0.75
Linear coefficient	0.42	0.30	1	1	0.50	0.48	0.93	0.96	1
Intercept t-estimate	-13.1	-11.6	0.18	0.11	-20.3	-19.9	0.39	0.37	0.93
Linear t-estimate	18.4	15.2	0.92	0.95	28.2	27.7	0.66	0.69	0.98
Linear regression Y_2 on Y_1									
Intercept coefficient	2.10	2.78	0.00	0.00	2.38	2.49	0.02	0.01	0.98
Linear coefficient	0.60	0.61	0.45	0.46	0.89	0.91	0.32	0.25	0.40
Intercept t-estimate	36.5	38.8	0.03	0.02	43.5	43.7	0.44	0.40	0.99
Linear t-estimate	17.6	14.0	0.95	0.97	28.2	27.7	0.66	0.69	0.96
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-0.80	-0.85	0.65	0.63	-1.07	-1.04	0.32	0.27	0.47
Linear coefficient	0.30	0.31	0.45	0.39	0.35	0.29	0.88	0.93	0.03
Quadratic coefficient	0.02	0.00	0.98	1	0.02	0.03	0.21	0.14	0.99
Intercept t-estimate	-10.6	-9.80	0.31	0.26	-16.1	-15.4	0.30	0.25	0.47
Linear t-estimate	5.46	5.42	0.50	0.44	7.44	6.26	0.86	0.91	0.03
Quadratic t-estimate	2.34	-0.22	0.98	0.44	3.28	4.07	0.18	0.12	0.98
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.48	3.38	0.00	0.00	1.17	1.30	0.05	0.03	1
Linear coefficient	0.60	0.61	0.44	0.43	0.89	0.91	0.32	0.25	0.43
Quadratic coefficient	0.21	-0.20	1	1	0.40	0.40	0.57	0.58	1
Intercept t-estimate	18.2	33.4	0.00	0.00	18.4	18.9	0.31	0.26	1
Linear t-estimate	19.4	15.8	0.95	0.99	36.4	34.6	0.75	0.83	0.99
Quadratic t-estimate	10.4	-7.85	1	1	25.2	23.5	0.79	0.88	1

Table 4.7: Simulation results for mean completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the Random Indicator method. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y_{com}^{rep})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y_{com}^{rep})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	0.37	0.25	0.81	0.83	2.16	2.15	0.51	1	1
Var of Y_2	1.72	1.60	0.73	0.70	5.69	5.64	0.52	1	1
Linear regression Y_1 on Y_2									
Intercept coefficient	-0.06	-0.02	0.45	0.51	-1.13	-1.13	0.48	1	0.84
linear coefficient	0.02	0.12	0.20	0.17	0.52	0.53	0.36	1	0.99
Intercept t-estimate	-0.89	-0.22	0.45	0.53	-21.0	-20.8	0.42	1	0.61
linear t-estimate	0.37	2.99	0.20	0.18	31.6	32.1	0.47	1	1
Linear regression Y_2 on Y_1									
Intercept coefficient	0.37	0.25	0.81	0.83	2.16	2.15	0.51	1	1
linear coefficient	0.01	0.06	0.20	0.18	0.97	0.97	0.49	1	1
Intercept t-estimate	9.75	6.23	0.81	0.83	40.8	41.4	0.51	1	1
linear t-estimate	0.54	3.20	0.19	0.17	31.5	31.9	0.47	1	1
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-0.03	0.00	0.41	0.41	-1.04	-1.04	0.40	1	0.66
Linear coefficient	-0.02	0.09	0.16	0.11	0.37	.38	0.42	1	1
Quadratic coefficient	-0.01	-0.01	0.53	0.69	0.03	0.02	0.57	1	0.56
Intercept t-estimate	-0.44	-0.00	0.40	0.41	17.84	-17.04	0.35	1	0.47
Linear t-estimate	-0.09	2.52	0.14	0.11	9.54	10.01	0.46	1	1
Quadratic t-estimate	-0.62	-0.48	0.50	0.68	3.84	3.45	0.61	1	0.40
Quadratic regression Y_2 on Y_1									
Intercept coefficient	0.27	0.24	0.60	0.61	0.66	0.67	0.51	1	0.90
Linear coefficient	0.01	0.06	0.20	0.18	0.96	0.96	0.49	1	1
Quadratic coefficient	0.03	0.00	0.92	1.00	0.49	0.49	0.50	1	1
Intercept t-estimate	4.78	3.94	0.59	0.63	14.2	14.6	0.51	0.90	
Linear t-estimate	0.53	3.20	0.20	0.17	53.3	55.3	0.38	1	1
Quadratic t-estimate	2.31	0.29	0.92	0.99	42.2	43.8	0.31	1	1

Table 4.8: Simulation results for median completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the Random indicator method. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	0.37	0.26	0.82	0.83	2.16	2.15	0.51	1	1
Var of Y_2	1.73	1.59	0.74	0.71	5.61	5.55	0.52	1	1
Linear regression Y_1 on Y_2									
Intercept coefficient	-0.05	-0.01	0.44	0.53	-1.12	-1.14	0.49	1	0.84
Linear coefficient	0.03	0.15	0.19	0.17	0.52	0.53	0.35	1	0.99
Intercept t-estimate	-0.76	-0.01	0.44	0.54	-20.9	-20.8	0.42	1	0.61
Linear t-estimate	0.63	3.35	0.20	0.17	31.7	32.1	0.47	1	1
Linear regression Y_2 on Y_1									
Intercept coefficient	0.38	0.26	0.81	0.83	2.15	2.15	0.51	1	1
Linear coefficient	0.01	0.06	0.20	0.17	0.97	0.97	0.50	1	1
Intercept t-estimate	9.85	6.52	0.82	0.83	40.7	41.4	0.51	1	1
Linear t-estimate	0.83	3.50	0.19	0.17	31.6	31.9	0.47	1	1
Quadratic regression Y_1 on Y_2									
Intercept coefficient	-0.02	0.02	0.38	0.39	-1.04	-1.03	0.40	1	0.66
Linear coefficient	0.00	0.12	0.14	0.09	0.37	0.38	0.41	1	1
Quadratic coefficient	-0.01	-0.01	0.53	0.73	0.02	0.02	0.58	1	0.56
Intercept t-estimate	-0.32	0.3	0.38	0.39	-17.7	-16.9	0.34	1	0.47
Linear t-estimate	0.15	2.78	0.13	0.10	9.6	10.0	0.47	1	1
Quadratic t-estimate	-0.59	-0.48	0.51	0.70	3.80	3.46	0.61	1	0.40
Quadratic regression Y_2 on Y_1									
Intercept coefficient	0.27	0.24	0.60	0.62	0.67	0.67	0.51	1	0.90
Linear coefficient	0.01	0.06	0.20	0.17	0.95	0.96	0.50	1	1
Quadratic coefficient	0.03	0.00	0.92	1.00	0.49	0.49	0.50	1	1
Intercept t-estimate	4.79	4.07	0.60	0.63	14.3	14.52	0.50	1	0.90
Linear t-estimate	0.82	3.51	0.19	0.17	52.7	54.8	0.38	1	1
Quadratic t-estimate	2.31	0.28	0.93	1	42.0	43.5	0.31	1	1

Table 4.9: Simulation results for mean completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = 1$ implying a weak MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y_{com}^{rep})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y_{com}^{rep})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	1.08	1.13	0.24	0.14	1.28	1.29	0.44	0.37	0.98
Var of Y_2	2.09	2.20	0.30	0.22	2.51	2.51	0.48	0.46	0.91
Linear regression Y_1 on Y_2									
Intercept coefficient	0.05	0.09	0.25	0.13	-0.28	-0.25	0.34	0.16	0.80
Linear coefficient	0.02	0.12	0.20	0.17	0.20	0.18	0.68	0.84	0.67
Intercept t-estimate	0.76	1.29	0.5	0.13	-4.39	-4.06	0.33	0.16	0.78
Linear t-estimate	-1.37	-2.22	0.74	0.87	6.78	6.25	.68	0.84	0.74
Linear regression Y_2 on Y_1									
Intercept coefficient	1.07	1.12	0.25	0.14	1.29	1.30	0.45	0.38	0.98
Linear coefficient	-0.03	-0.06	0.75	0.88	0.20	0.19	0.67	0.83	0.77
Intercept t-estimate	0.76	1.29	0.25	0.13	27.1	27.1	0.48	0.45	0.98
Linear t-estimate	-1.37	-2.22	0.74	0.88	6.24	5.76	0.66	0.80	0.76
Quadratic regression Y_1 on Y_2									
Intercept coefficient	0.05	0.09	0.26	0.13	-0.24	-0.22	0.34	0.50	0.79
Linear coefficient	-0.07	-0.05	0.44	0.39	0.01	0.01	0.50	0.50	0.64
Quadratic coefficient	0.00	-0.02	0.73	0.89	0.05	0.04	0.59	0.73	0.84
Intercept t-estimate	0.81	1.33	0.26	0.13	-3.68	-3.31	0.33	0.18	0.75
Linear t-estimate	-1.05	-0.81	0.43	0.38	0.25	0.25	050	0.50	0.64
Quadratic t-estimate	0.23	-0.59	0.73	0.89	3.40	3.06	0.60	0.73	0.85
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.08	1.22	0.09	0.01	1.00	1.00	0.51	0.51	1
Linear coefficient	-0.03	-0.06	0.75	0.88	0.20	0.19	0.67	0.83	0.80
Quadratic coefficient	0.00	-0.03	0.91	1.00	0.10	0.10	0.44	0.40	1
Intercept t-estimate	15.8	17.6	0.12	0.01	14.5	14.4	0.52	0.53	1
Linear t-estimate	-1.36	-2.20	0.74	0.88	7.10	6.54	0.67	0.84	0.77
Quadratic t-estimate	0.53	-1.30	0.91	0.99	5.46	5.59	0.45	0.40	1

Table 4.10: Simulation results for median completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = 1$ implying a weak MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	1.05	1.09	0.25	0.14	1.27	1.28	0.44	0.37	0.98
Var of Y_2	1.94	2.02	0.29	0.21	2.40	2.40	0.48	0.47	0.91
Linear regression Y_1 on Y_2									
Intercept	0.05	0.09	0.24	0.11	-0.27	-0.24	0.34	0.15	0.80
Linear coefficient	-0.06	-0.09	0.74	0.90	0.20	0.18	0.68	0.85	0.67
Intercept t-estimate	0.82	1.34	0.24	0.10	-4.40	-4.05	0.33	0.84	0.78
Linear t-estimate	-1.51	-2.36	0.76	0.90	6.78	6.25	0.68	0.84	0.74
Linear regression Y_2 on Y_1									
Intercept coefficient	1.05	1.08	0.26	0.15	1.28	1.29	0.45	0.38	0.98
Linear coefficient	-0.04	-0.06	0.76	0.90	0.18	0.17	0.67	0.84	0.77
Intercept t-estimate	24.4	24.9	0.34	0.88	27.7	27.9	0.47	0.43	0.98
Linear t-estimate	-1.40	-2.26	0.74	0.88	6.74	6.41	0.67	0.82	0.76
Quadratic regression Y_1 on Y_2									
Intercept coefficient	0.05	0.10	0.26	0.11	-0.24	-0.21	0.34	0.16	0.79
Linear coefficient	-0.06	-0.05	0.44	0.37	0.01	0.01	0.51	0.50	0.64
Quadratic coefficient	0.00	-0.02	0.73	0.93	0.05	0.04	0.59	0.75	0.84
Intercept t-estimate	0.83	1.38	0.26	0.11	-3.60	-3.23	0.33	0.17	0.75
Linear t-estimate	-1.03	-0.80	0.43	0.11	0.15	0.17	0.50	0.51	0.64
Quadratic t-estimate	0.15	-0.72	0.74	0.93	3.73	3.39	0.60	0.74	0.85
Quadratic regression Y_2 on Y_1									
Intercept coefficient	1.00	1.11	0.09	0.00	0.99	0.99	0.51	0.51	1
Linear coefficient	-0.04	-0.06	0.77	0.90	0.18	0.17	0.67	0.84	0.80
Quadratic coefficient	0.02	-0.01	0.91	1.00	0.08	0.08	0.44	0.38	1
Intercept t-estimate	15.4	17.1	0.11	0.01	14.4	14.4	0.52	0.53	1
Linear t-estimate	-1.49	-2.36	0.75	0.90	6.95	6.51	0.67	0.85	0.77
Quadratic t-estimate	1.15	-0.55	0.91	0.99	4.90	5.10	0.45	0.39	1

Table 4.11: Simulation results for mean completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = \infty$ implying a strong MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	0.96	0.98	0.41	0.36	1.23	1.25	0.33	0.19	0.23
Var of Y_2	1.72	1.77	0.44	0.42	2.43	2.47	0.38	0.28	0.25
Linear regression Y_1 on Y_2									
Intercept coefficient	0.02	0.04	0.37	0.31	-0.14	-0.10	0.24	0.09	0.19
Linear coefficient	-0.02	-0.04	0.63	0.70	0.09	0.06	0.79	0.91	0.15
Intercept t-estimate	0.32	0.63	0.37	0.31	-2.09	-1.52	0.24	0.09	0.17
Linear t-estimate	-0.63	-1.13	0.63	0.70	3.24	2.33	0.77	0.91	0.15
Linear regression Y_2 on Y_1									
Intercept coefficient	0.96	0.97	0.41	0.36	1.24	1.26	0.33	0.20	0.25
Linear coefficient	-0.01	-0.03	0.63	0.70	0.11	0.08	0.77	0.91	0.18
Intercept t-estimate	23.4	23.6	0.44	0.41	26.0	26.2	0.42	0.32	0.35
Linear t-estimate	-0.60	-1.06	0.62	0.68	3.23	2.33	0.77	0.91	0.21
Quadratic regression Y_1 on Y_2									
Intercept coefficient	0.02	0.04	0.36	0.27	-0.11	-0.07	0.24	0.09	0.86
Linear coefficient	-0.04	-0.01	0.37	0.22	-0.02	-0.01	0.47	0.46	0.17
Quadratic coefficient	0.01	-0.01	0.75	0.91	0.02	0.01	0.68	0.82	0.76
Intercept t-estimate	0.27	0.59	0.36	0.27	-1.56	-0.97	0.24	0.09	0.16
Linear t-estimate	-0.66	-0.17	0.36	0.22	-0.25	-0.16	0.48	0.46	0.76
Quadratic t-estimate	0.39	-0.56	0.75	0.91	1.97	1.30	0.68	0.81	0.62
Quadratic regression Y_2 on Y_1									
Intercept coefficient	0.87	0.96	0.15	0.01	1.04	1.06	0.43	0.36	0.35
Linear coefficient	-0.02	-0.03	0.63	0.70	0.11	0.08	0.76	0.91	1
Quadratic coefficient	0.03	0.00	0.88	0.99	0.07	0.07	0.44	0.39	0.20
Intercept t-estimate	14.2	15.5	0.17	0.02	14.9	14.9	0.48	0.45	1
Linear t-estimate	-0.62	-1.13	0.63	0.70	3.30	2.39	0.77	0.91	0.16
Quadratic t-estimate	1.96	0.30	0.88	0.99	3.67	3.81	0.45	0.40	1

Table 4.12: Simulation results for median completed-data discrepancy with $n = 1000$, with 80% of missingness on y_2 when the missing values are imputed using the proxy pattern mixture hot deck assuming the sensitivity parameter $\lambda = \infty$ implying a strong MNAR mechanism. Imputations are made assuming a correctly specified quadratic relationship of y_2 on y_1 and an incorrectly specified linear relationship of y_2 on y_1. We also present the proportion (Prop) of times the posterior predictive p-value is closer to 0.5 for the correctly specified imputation model

	Linear Imputation Model				Quadratic Imputation Model				Prop
Analysis	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	$Q(\bar{Y}_{obs}, Y_{mis})$	$\bar{Q}(Y^{rep}_{com})$	$p_{B,com}$	$p_{B,ecom}$	
Mean of Y_2	0.94	0.96	0.44	0.38	1.22	1.23	0.32	0.16	0.23
Var of Y_2	1.66	1.68	0.46	0.44	2.29	2.36	0.38	0.24	0.25
Linear regression Y_1 on Y_2									
Intercept coefficient	0.02	0.04	0.38	0.28	-0.13	-0.09	0.24	0.06	0.19
Linear coefficient	-0.02	-0.04	0.61	0.72	0.09	0.06	0.79	0.94	0.15
Intercept t-estimate	0.33	0.64	0.38	0.28	-1.81	-1.22	0.23	0.06	0.17
Linear t-estimate	-0.55	-0.91	0.62	0.73	2.86	1.94	0.78	0.94	0.15
Linear regression Y_2 on Y_1									
Intercept coefficient	0.94	0.96	0.44	0.38	1.22	1.25	0.33	0.18	0.25
Linear coefficient	-0.01	-0.02	0.62	0.73	0.08	0.06	0.77	0.94	0.18
Intercept t-estimate	23.6	23.8	0.45	0.43	26.1	26.4	0.42	0.31	0.35
Linear t-estimate	-0.59	-0.82	0.58	0.65	2.87	1.94	0.78	0.94	0.21
Quadratic regression Y_1 on Y_2									
Intercept coefficient	0.02	0.01	0.38	0.25	-0.10	-0.06	0.23	0.07	0.86
Linear coefficient	-0.04	-0.01	0.36	0.20	-0.02	-0.01	0.48	0.45	0.17
Quadratic coefficient	0.10	-0.01	0.76	0.96	0.03	0.02	0.67	0.86	0.76
Intercept t-estimate	0.229	0.60	0.38	0.25	-1.45	-0.82	0.23	0.07	0.16
Linear t-estimate	-0.65	-0.13	0.36	0.20	-0.23	-0.11	0.48	0.46	0.76
Quadratic t-estimate	0.39	-0.52	0.76	0.97	2.19	1.56	0.67	0.86	0.62
Quadratic regression Y_2 on Y_1									
Intercept coefficient	0.86	0.94	0.15	0.01	1.04	1.05	0.43	0.36	0.35
Linear coefficient	-0.01	-0.02	0.62	0.72	0.08	0.06	0.77	0.94	1
Quadratic coefficient	0.03	0.01	0.89	1.00	0.06	0.07	0.42	0.35	0.20
Intercept t-estimate	14.24	15.5	0.16	0.01	14.7	14.8	0.48	0.45	1
Linear t-estimate	-0.55	-0.93	0.62	0.72	2.82	1.93	0.78	0.94	0.16
Quadratic t-estimate	2.05	0.36	0.89	0.99	3.49	3.63	0.44	0.37	1

Table 4.13: Diagnostic results for the mean National Institute of Health Stroke Score under the imputation models for the GCASR registry data

	Correct Model			Misspecified Model		
Method	$Q(Y_{obs}, Y_{mis})$	$Q(Y^{rep}_{com})$	$p_{B,com}$	$Q(Y_{obs}, Y_{mis})$	$Q(Y^{rep}_{com})$	$p_{B,com}$
RI	6.71	6.71	0.50	7.04	6.93	1
PPMHD(1)	6.91	6.37	1	7.37	7.51	0.36
PPMHD(∞)	6.59	4.90	0.99	7.37	7.33	0.70
NBI(3,0.2,0.8)	7.20	7.24	0.56	7.21	7.30	0.34
NBI(3,0.5,0.5)	6.92	6.98	0.53	7.46	7.35	0.92
NBI(3,0.8,0.2)	6.99	7.06	0.55	7.54	7.48	0.74
NMI(3,0.2,0.8)	7.04	6.96	0.68	7.49	7.49	0.51
NMI(3,0.5,0.5)	7.06	6.99	0.79	7.43	7.71	0
NMI(3,0.8,0.2)	7.09	6.97	0.86	7.51	7.96	0

Chapter 5

Future Work

There are numerous extensions that can arise from this dissertation. As described in Chapter 2, we perform multiple imputation (MI) using dimension reduction techniques for high-dimensional to impute one variable with missing data. Also in Chapter 3, we handle non-ignorable missingness by using a nonparametric imputation method that harnesses the power of bootstrap imputation and multiple imputation. We use our nonparametric imputation method to impute a single variable with missing data. However, in practice there are often multiple variables with missing data with a general missing data pattern. Although methods for general missing data patterns have been suggested (Deng et al., 2016; Zhao and Long, 2013b), its theoretical properties are not well-established. Future development to extend imputation methods to handle general missing data patterns with more rigorous theoretical justification is beneficial.

Some other extensions of this dissertation pertain to nonparametric imputation. In Chapter 3, we describe a nonparametric method for handling non-ignorable missing data. Thus far, we focus on estimation of the mean. A useful extensions would be to include regression estimation in the subsequent analysis. With regression estimation, we could focus on potential bias and the nominal coverage for parameter estimates.

In Chapter 4, we evaluate posterior predictive checking for diagnosing problems with imputation models assuming non-ignorable nonresponse. We consider continuous data with missing values in simulations. The next obvious step is to determine the appropriateness of diagnostics for identifying discrepancies in imputation when the data are binary or categorical. In addition, when a single variable has missing values, application of posterior predictive checking is straightforward. The more complex case is when there are multiple variables with missing values. A logical extension is to develop a diagnostic method for imputation models for general missing data patterns.

Bibliography

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

Bondarenko, I. and Raghunathan, T. (2016), 'Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models', *Statistics in Medicine* .

Camp, D., Bryant, K., Zimmermann, S., Brasher, C., Connelly, K. M., Dunn, J., Frankel, M., Ido, M., Lugtu, J. and Nahab, F. (2015), Presenting symptoms and response to dysphagia screen predict unfavorable outcome in acute ischemic stroke patients who do not receive iv tpa due to mild and rapidly improving stroke symptoms, *in* 'Stroke', Vol. 46.

Carpenter, J. and Kenward, M. (2012), *Multiple imputation and its application*, John Wiley & Sons.

Carpenter, J. R., Kenward, M. G. and White, I. R. (2007), 'Sensitivity analysis after multiple imputation under missing at random: a weighting approach.', *Statistical methods in medical research* **16**, 259–275.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Deng, Y., Chang, C., Ido, M. S. and Long, Q. (2016), 'Multiple imputation for general missing data patterns in the presence of high-dimensional data', *Scientific reports* **6**.

Dennis Cook, R. (2000), 'Save: a method for dimension reduction and graphics in regression', *Communications in statistics-Theory and methods* **29**(9-10), 2109–2121.

Efron, B. (1994), 'Missing data, imputation, and the bootstrap', *Journal of the American Statistical Association* **89**(426), 463–475.

Fan, J., Feng, Y., Saldana, D. F., Samworth, R. and Wu, Y. (2014), *SIS: Sure Independence Screening*. R package version 0.7-4.

Fan, J. and Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Gelman, A., Meng, X.-L. and Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica sinica* **6**(4), 733–760.

Glynn, R. J., Laird, N. M. and Rubin, D. B. (1993), 'Multiple imputation in mixture models for nonignorable nonresponse with follow-ups', *Journal of the American Statistical Association* **88**(423), 984–993.

Harel, O. and Zhou, X.-H. (2007), 'Multiple imputation: review of theory, implementation and software', *Statistics in medicine* **26**(16), 3057–3077.

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

Heckman, J. J. (1977), 'Sample selection bias as a specification error (with an application to the estimation of labor supply functions)'.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161.

Jolani, S. (2012), 'Dual Imputation Strategies for Analyzing Incomplete Data'.

Kenward, M. G. and Molenberghs, G. (1998), 'Likelihood based frequentist inference when data are missing at random', *Statistical Science* pp. 236–247.

Kim, J. Y. and Kim, J. K. (2012), 'Parametric fractional imputation for nonignorable missing data', *Journal of the Korean Statistical Society* **41**(3), 291–303.

Lee, S., Epstein, M., Duncan, R. and Lin, X. (2012), 'Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome Wide Association Studies', *Genetic Epidemiology* **36**(4), 293–302.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**(414), 316–327.

Li, K.-C. (1992), 'On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma', *Journal of the American Statistical Association* **87**(420), 1025–1039.

Liao, S., Lin, Y., Kang, D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C. and Tseng, G. (2014), 'Missing value imputation in high-dimensional phenomic data: imputable or not, and how?', *BMC Bioinformatics* **15**(1), 1.

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

Little, R. J. (1988*b*), 'A test of missing completely at random for multivariate data with missing values', *Journal of the American Statistical Association* **83**(404), 1198–1202.

Little, R. J. (1993), 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association* **88**(421), 125–134.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Long, Q., Hsu, C.-H. and Li, Y. (2012), 'Doubly robust nonparametric multiple imputation for ignorable missing data', *Statistica Sinica* **22**, 149.

Meng, X.-L. (1994*a*), 'Multiple-imputation inferences with uncongenial sources of input', *Statistical Science* pp. 538–558.

Meng, X.-L. (1994*b*), 'Posterior predictive p-values', *The Annals of Statistics* pp. 1142–1160.

Nguyen, C. D., Carlin, J. B. and Lee, K. J. (2013), 'Diagnosing problems with imputation models using the kolmogorov-smirnov test: a simulation study', *BMC medical research methodology* **13**(1), 1.

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

Park, T. and Casella, G. (2008), 'The bayesian lasso', *Journal of the American Statistical Association* **103**(482), 681–686.

Poirier, D. J. and Ruud, P. A. (1983), 'Diagnostic testing in missing data models', *International Economic Review* pp. 537–546.

Raghunathan, T. and Bondarenko, I. (2007), 'Diagnostics for multiple imputations', *Available at SSRN 1031750* .

Richardson, J., Murray, D., House, C. K. and Lowenkopf, T. (2006), 'Successful implementation of the national institutes of health stroke scale on a stroke/neurovascular unit', *Journal of Neuroscience Nursing* **38**(4), 309.

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

Rubin, D. B. (1978), Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse, *in* 'Proceedings of the survey research methods section of the American Statistical Association', Vol. 1, American Statistical Association, pp. 20–34.

Rubin, D. B. (1987), 'Multiple imputation for nonresponse in surveys (wiley series in probability and statistics)'.

Sande, I. G. (1982), 'Imputation in surveys: coping with reality', *The American Statistician* **36**(3a), 145–152.

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

Schafer, J. P. S. U. (1999), 'Multiple Imputation: a primer', *Statistical methods in medical research* **8**, 3–15.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999), 'Adjusting for nonignorable drop-out using semiparametric nonresponse models', *Journal of the American Statistical Association* **94**(448), 1096–1120.

Siddique, J. and Belin, T. R. (2008), 'Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data.', *Computational statistics & data analysis* **53**(2), 405–415.

Siddique, J., Harel, O. and Crespi, C. M. (2012), 'Addressing Missing Data Mechanism Uncertainty using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial', *Annals of Applied Statistics* **6**(4), 1814–1837.

Stekhoven, D. J. and Bühlmann, P. (2012), 'Missforest-Non-parametric missing value imputation for mixed-type data', *Bioinformatics* **28**(1), 112–118.

Sullivan, D. and Andridge, R. (2015), ‘A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck’, *Computational Statistics & Data Analysis* **82**, 173–185.

Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011), ‘mice: Multivariate imputation by chained equations in r’, *Journal of Statistical Software* **45**(3), 1–67.

Wang, J. (2010), Diagnostics for multiple imputation based on the propensity score, PhD thesis.

Weisberg, S. (2002), ‘Dimension reduction regression in r’, *Journal of Statistical Software* **7**(1), 1–22.

White, I. R. and Carlin, J. B. (2010), ‘Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values’, *Statistics in medicine* **29**(28), 2920–2931.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009), ‘A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.’, *Biostatistics (Oxford, England)* **10**(3), 515–34.

Witten, D., Tibshirani, R., Gross, S. and Narasimhan, B. (2013), *PMA: Penalized Multivariate Analysis*. R package version 1.0.9.

Zhao, Y. and Long, Q. (2013*a*), ‘Multiple imputation in the presence of high-dimensional data.’, *Statistical methods in medical research* .

Zhao, Y. and Long, Q. (2013*b*), ‘Multiple imputation in the presence of high-dimensional data’, *Statistical methods in medical research* p. 0962280213511027.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2006), 'Sparse Principal Component Analysis', *Journal of Computational and Graphical Statistics* **15**(2), 265–286.