Distribution Agreement

Signature:

______________________________ ________________________

Yi Deng Date

Statistical Methods for Incomplete Big Data

By

Yi Deng

Doctor of Philosophy

Biostatistics

Qi Long, Ph.D.
Advisor

Howard Chang, Ph.D.
Committee Member

Xiaoqian Jiang, Ph.D.
Committee Member

Robert H. Lyles, Ph.D.
Committee Member

Accepted:

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

Date

Statistical Methods for Incomplete Big Data

By

Yi Deng

B.S., University of Science and Technology of China, 2013

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

2017

Abstract

Statistical Methods for Incomplete Big Data

By

Yi Deng

Advances in technology have led to generation of enormous amounts of data, also known as "big data". Such explosion in turn brings daunting challenges for data analysis and for generating meaningful findings using big data. One major challenge is the occurrence of missing data. Data insights may be impacted if missing values are inadequately handled. In this dissertation, we develop and investigate methods for handling missing data in the environment of big data.

In Chapter 1, we first review the terminology on missing data and existing methods for handling incomplete data or big data. Furthermore, we present distributed analyses of big data that are stored in multiple sources.

In Chapter 2, we develop two approaches of using regularized regressions to impute missing values in the presence of high-dimensional big data. The approaches can accommodate mixed incomplete data and handle general missing data patterns. Our approaches are compared to several existing imputation methods in simulation studies. The simulation results demonstrate that the proposed multiple imputation approach based on an indirect use of regularized regression outperforms any other imputation methods.

In addition to traditional types of data with missing values, this dissertation also investigates handling distributed incomplete data, with the purpose of protecting the privacy. For example, in the case of medical patients, institutions such as the Veteran's Health Administration have policies that restrict their data to internal facilities. Under such circumstances, distributed analyses are necessary but challenging when data are subject to missing values. In Chapter 3, we propose privacy-preserving methods to handle missing data in distributed analyses for horizontally partitioned data. The methods, in particular, target data that are missing at random and missing not at random. In Chapter 4, we present privacy-preserving methods on vertically partitioned data with missing values.

Statistical Methods for Incomplete Big Data

By

Yi Deng

B.S., University of Science and Technology of China, 2013

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
2017

Acknowledgement

Foremost, I would like to thank my advisor Dr. Qi Long, who introduced me to the biostatistics research and provided continuous support for my Ph.D study. His immense knowledge, enormous enthusiasm, and other qualities make him my role model that will affect my future career and life. Without his invaluable help and guidance, this dissertation would never be possible.

Besides my advisor, I would like to express my deepest appreciation to my committee members: Dr. Howard Chang, Dr. Xiaoqian Jiang and Dr. Robert H. Lyles for generously offering their time and guidance throughout the preparation and review of this dissertation. Their suggestions have significantly improved the work and my presentation skill.

I would also like to thank Dr. Lawrence S. Phillips, Dr. Mary Rhee, Dr. Yi-An Ko, Dr. Sandra Safo and other faculty members that I have collaborated with. The experiences are enjoyable and I have acquired plenty knowledge by learning from them. My sincere thanks also go to my Emory colleagues and friends, for accompanying me throughout this wonderful Ph.D journey, and for all the memories we have had in the last four years.

Last but not least, I would like to thank my parents for their support throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The world has been generating huge volumes of data from all aspects of our lives at unprecedented high speed. Such large amounts of data (also known as “big data”) provide a wealth of opportunities by deriving hidden insights for a better decision making. For example, big data in healthcare, exploded by rapid digitization of prescriptions, medical images, laboratory results, insurance information and etc., are a key to improve the quality of healthcare delivery meanwhile reducing the costs (Raghupathi and Raghupathi, 2014).

Although “big” data is actually a relative term to some extent, it generally refers to large and complex data that are difficult to manage and analyze with traditional strategies. To address the challenge of the management of big data, people usually adopt the distributed file system such as Hadoop that allows the data to store and read in parallel across nodes in a cluster. The analytical challenge of big data arises as traditional statistical analyses require the entire dataset to be loaded in memory before applying any methods. The algorithms (e.g., functions written in R) assuming the pooled data can fit in the RAM on a single computer are no longer applicable and thus can not be deployed on a distributed system when the data are too large. Even if engineers invent a RAM that is large enough for pooled big data, the process of integrating/pooling the data has severe privacy issues, as we describe in Chapter 3 and 4. To address this, researchers propose distributed algorithms for models such as linear regression and logistic regression, that would work well in a distributed environment.

While the beauty of statistics is to reveal valuable and reliable latent information from data in a wide variety of areas, this nature is deteriorated by the presence of missing data, a challenge that lies in data of any volumes. By definition, missing Data are underlying observations that exist but are not stored for some reason, and are ubiquitous in biomedical research. For example, in a clinical survey study, missing data often result from subjects’ non-responses of all or some of the questions in the questionnaire. In a longitudinal study where each observational unit is measured repeatedly over a period of time, subjects may drop out or be unable to participate at all time points. In an analysis with missing data, a relatively few absent observations can dramatically shrink the sample size, leading to inaccurate statistical inference, weaker statistical power and perhaps biased estimates. Therefore, handling missing data is essential. This dissertation studies methods for missing data with a focus on incomplete big data. In the remainder of this chapter we first describe missing-data patterns and mechanisms, which need to be examined properly before any techniques to be applied. We then consider a literature review of conventional missing data methods and distributed analysis. Chapter 2 is concerned with multiple imputation methods for high-dimensional data with general missing data patterns. Chapter 3 and Chapter 4 study privacy-preserving methods for handling missing data in a distributed environment where data are either horizontally or vertically partitioned. For all of the methods that we propose, extensive simulation studies are conducted to evaluate the performances and the data collected by the Georgia Coverdell Acute Stroke Registry (GCASR) are used as a motivation and example. We conclude this dissertation with a summary and future work in Chapter 5.

1.1 Missing-Data Patterns and Mechanisms

Missing-data patterns define the type of missingness depending on values observed and missing in the data set. Missing-data mechanisms capture the relationship between the missingness of values of certain variables and values of all variables in the data set. Let $Z = (z_{ij})$ be an $n \times p$ data matrix with i-th row a p-vector observation of variables $(Z_1, ..., Z_p)$ for subject i. In the presence of missing data, we further let $R = (r_{ij})$ be the matrix of binary missing-data indicators of the same dimension as the data set Z, with $r_{ij} = 1$ if z_{ij} is observed and $r_{ij} = 0$ otherwise. If the missing values only occur on one variable but a set of the rest variables is fully observed, we call it a univariate missing-data pattern. Some methods are restricted to univariate missing-data patterns since such patterns are simple and the methods are easy to illustrate. A data set has a monotone missing-data pattern if its variables could be arranged so that if Z_j is missing for a subject, then $Z_{j+1}, ..., Z_p$ are missing as well. The monotone missing-data pattern often arises in longitudinal studies that suffer from subjects' dropouts. A more common missing-data pattern is the general pattern. A data set has a general missing-data pattern when any set of variables may be missing for any subjects. Figure 1.1 illustrates the three missing-data patterns.

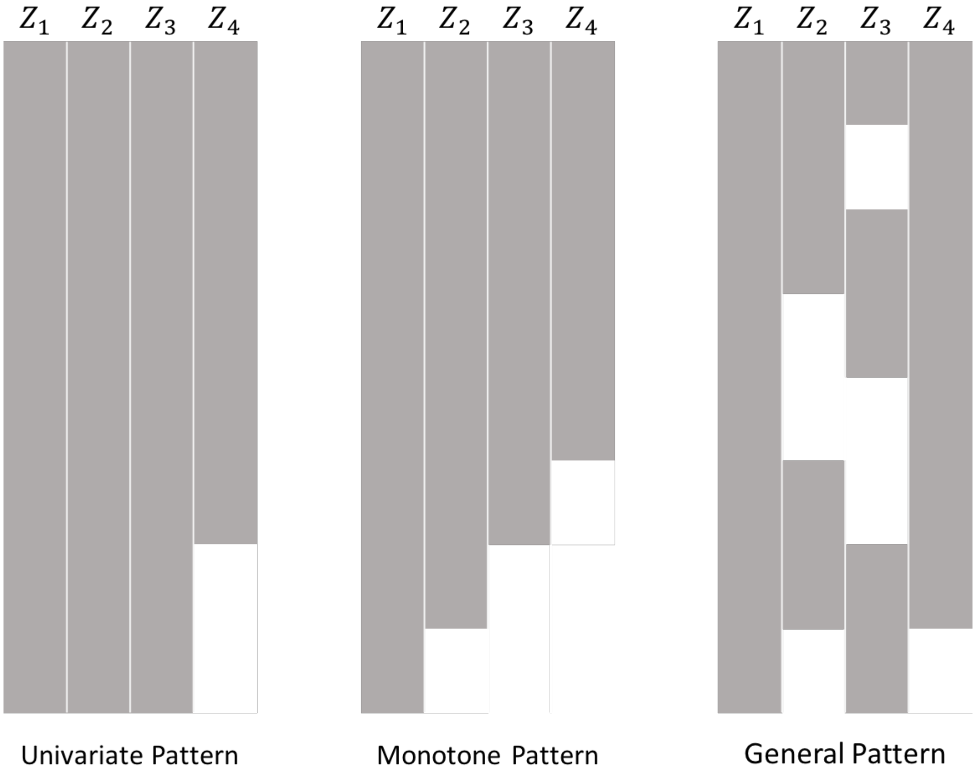


Figure 1.1: Visual illustrations of three missing-data patterns with 4 variables. Missing values are represented by unshaded areas

Besides the missing-data pattern, another key concept in missing data is the mechanism that leads to missing data. By introducing random missing-data indicators as we defined above, Rubin (1976) first systematically formalizes the missing-data mechanisms via probability distribution. The taxonomy of missing-data mechanisms according to the aforementioned paper, which distinguishes between missing completely at random

(MCAR), missing at random (MAR) and missing not at random (MNAR), is universally used. According to the framework proposed by Rubin (1987); Little and Rubin (2002), the missing-data mechanism is characterized by the conditional distribution of R given Z, i.e., $f(R|Z,\phi)$, where ϕ represents unknown parameters. The strongest assumption made for the data are MCAR, in which the distribution of R is independent of the values of the data Z, that is, $f(R|Z,\phi) = f(R|\phi)$ for all Z, ϕ, which implies $f(Z|R,\phi) = f(Z|\phi)$. In this situation, the observed data set can be considered as a random subset of the complete data set. Let Z_{obs} be the observed components of Z and Z_{mis} the missing components. A less restrictive assumption for data Z is that the probability of missing data on Z depends only on Z_{obs} but not on Z_{mis}. In this case, we call the data are MAR. That is, $f(R|Z,\phi) = f(R|Z_{obs},\phi)$ for all Z_{mis}, ϕ, which implies $f(Z|Z_{obs},R,\phi) = f(Z|Z_{obs},\phi)$. Unlike MCAR, when data are MAR, the observed data set is no longer a simple subset of the complete data set. On the other hand, MAR mechanism "requires that the missing values behave like a random sample of all values within subclasses defined by observed data" (Schafer, 1997). When the distribution of R depends on the missing component Z_{mis}, we call the data are MNAR. It is an important area of research that is yet to be thoroughly investigated.

Although we can determine the missing-data pattern by observed and missing values, there is no test that can definitively examine whether the missing-data mechanism is MCAR, MAR or MNAR. Thus, methods for handling missing data depend on the missing-data pattern and our assumption of the missing-data mechanism. Next, we discuss some commonly-used missing data methods with their corresponding assumptions of the mechanisms.

1.2 Commonly-Used Missing Data Methods

In this section, we briefly review some commonly-used methods for handling missing data. We call a subject a respondent if it does not have any missing values and call it non-respondent otherwise. To start with, we introduce general criteria for evaluating a missing-data method, such criteria are illustrated in Millsap and Maydeu-Olivares (2009, Chapter 5). Specifically, a superior method is expected to achieve the following qualities: 1. Minimized bias. Even though missing data may yield biased estimates, the good method is expected to make the bias as small as possible; 2. Sufficient use of the available information. The good method should not discard any pieces of data; 3. Good estimation of uncertainty. Particularly, we desire the method could achieve accurate estimates of standard errors, confidence intervals and p-values.

1.2.1 Complete-Case Analysis

Standard statistical methods for regression analysis have been widely developed to analyze rectangular data sets, where values of all variables are measured for all subjects. Therefore, in the presence of missing data, a natural solution is to discard those subjects with any missing values and conduct the analysis based on the subset of complete cases. The aforementioned method, called complete-case (CC) analysis or list-wise deletion, is a default procedure for handling missing data in many statistical packages such as R and SAS. There is a major advantage to CC analysis: it can be directly used for any kind of statistical analysis (e.g., survival analysis, longitudinal analysis) without any computational methods for dealing with missing data (Allison, 2001). CC analysis obtains some attractive statistical properties as well, relying on the missing-data mechanism. If the data are MCAR, CC analysis would provide unbiased estimates since the reduced sample of fully observed subjects is a random sub-sample of the original sample. However, the standard errors from the CC analysis will usually be larger since a sub-sample is utilized, leading to a wider confidence interval. On the other hand, if the data are MAR or MNAR, CC analysis will generally lead to biased results with few exceptions as have been discussed in detail by Rubin (1987); Little and Rubin (2002). For instance, the model of generalized linear regression using CC when the missingness only depends on independent variable, provides unbiased coefficients estimates. Another case that CC analysis is valid is when we fit a logistic regression using incomplete data and the probability of missing values on any variable depends only on the value of the dependent variable but not on any of the independent variables. In that case, the CC analysis yields consistent estimates of the regression coefficients and their standard errors (Vach, 2012).

Table 1.1 summarizes whether we expect biased estimates from a CC analysis in the case of linear regression and logistic regression, assuming different missing-data mechanisms. For both types of regression, Y is the outcome and $(X,\ Z)$ are covariates, where Z are fully observed. The table shows that the validity of CC analysis depends not on which variable is missing but on the mechanism leading to missingness.

Table 1.1: Validity of complete-case analysis in linear regression and logistic regression. Coef., coefficient estimates

Variable missing	Mechanism depends on	Linear regression		Logistic regression	
		Coef. of X	Coef. of Z	Coef. of X	Coef. of Z
Y	Y (MNAR)	Biased	Biased	Unbiased	Unbiased
	X, Z (MAR)	Unbiased	Unbiased	Unbiased	Unbiased
	Y, X (MNAR)	Biased	Biased	Biased	Unbiased
	Y, Z (MNAR)	Biased	Biased	Unbiased	Biased
	Y, X, Z (MNAR)	Biased	Biased	Biased	Biased
X	Y (MAR)	Biased	Biased	Unbiased	Unbiased
	X, Z (MNAR)	Unbiased	Unbiased	Unbiased	Unbiased
	Y, X (MNAR)	Biased	Biased	Biased	Unbiased
	Y, Z (MAR)	Biased	Biased	Unbiased	Biased
	Y, X, Z (MNAR)	Biased	Biased	Biased	Biased

1.2.2 Single and Multiple Imputation

In order to retain the sample size of a data set for review in the presence of missing data, a conventional strategy named single imputation is commonly implemented. Specifically, the strategy substitutes each missing value with a reasonable guess or estimate. One simple but popular single-imputation approach is to fill in the missing value with the unconditional mean for the cases that observe the variable. This well-known approach is studied in Haitovsky (1968) and is shown to bias the estimates. Unlike the unconditional mean substitution, regression imputation predicts the most likely value of missing data using a regression model fitted by the observed values of a variable on other variables. Another ad hoc single imputation method is called hot-deck imputation which replaces each missing value with a random draw from the observed values. One comparatively convenient hot-deck approach used in longitudinal studies is last observation carried forward (LOCF). It sorts the data matrix based on any set of variables and then fills in the missing value with the closest observed value ahead. LOCF is built upon the belief that, for example in a longitudinal study with repeated measurements, the missing measurement does not change from the last time it is measured/observed. Little and Rubin (2002) and Allison (2001) show that single-imputation methods tend to underestimate standard errors since the methods cannot distinguish real data from imputed data and are unable to justify the uncertainty in the imputations.

By contrast, multiple imputation (MI) method replaces each missing value with a set of M plausible values. The method is considered as an improvement upon single imputation. Generally, MI procedure consists three steps: 1. imputation step, in which the missing data are imputed and M complete data sets are generated; 2. analysis step, in which a standard statistical technique is applied to analyze each complete data set; 3. combining step, in which the results of the above analyses are combined to provide a final result that accounts for the uncertainty in the data as well as that due to missing values. Figure 1.2 is a pictorial representation of these three steps in a general MI method. We further delve deeply into each step, with introductions of some additional terminologies for MI in the meantime.

Incomplete data are considered to have ignorability if they meet the following two con-

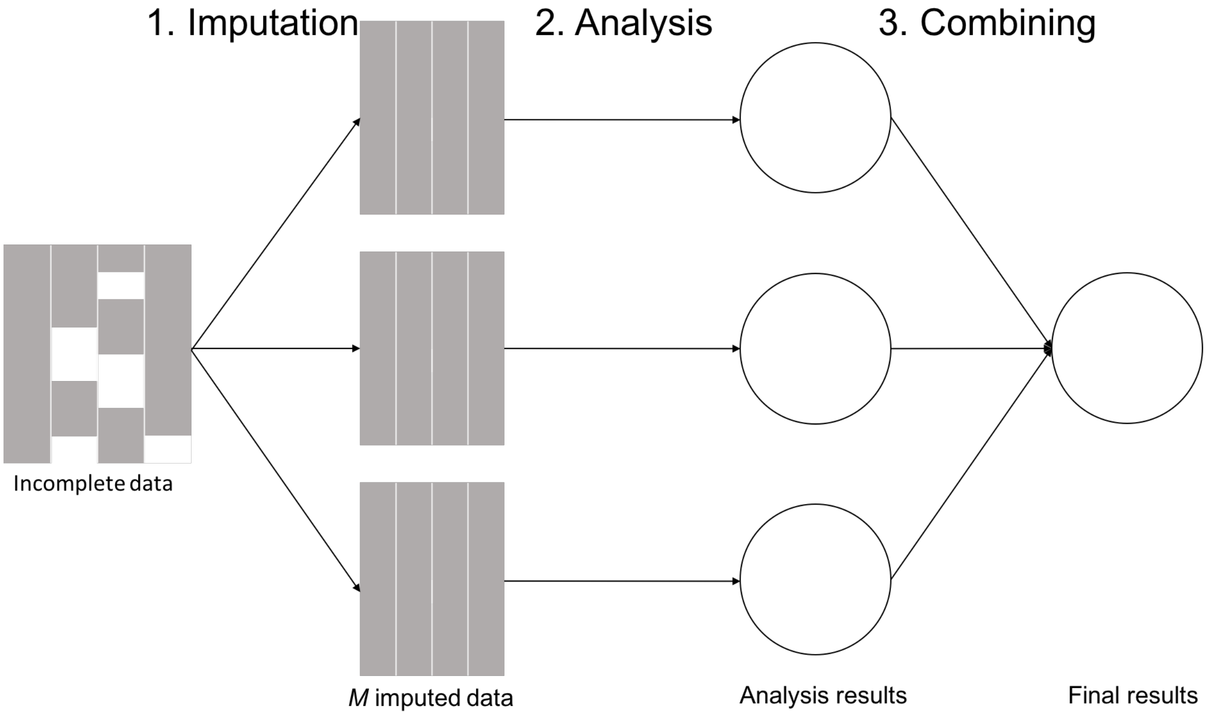


Figure 1.2: Pictorial representation of three steps in multiple imputation methods

ditions: 1. MAR; 2. parameters governing the missingness are distinct from parameters to be estimated. To make it clear, let's consider a Bayesian joint model for the complete data and the missingness with underlying parameters that govern the data

$$f(Z, R, \theta, \phi) = f(Z|\theta)f(R|Z, \phi)f(\theta, \phi) \tag{1.1}$$

where θ is the unknown parameters of the complete data that we want to estimate and ϕ is from the conditional distribution of missing indicators R given the complete data Z. Ignorability requires the data to be MAR and that the joint prior distribution for θ and ϕ can factor into separate parts: $f(\theta, \phi) = f(\theta)f(\phi)$. Under ignorability, we can show that $f(Z_{mis}|Z_{obs}, R) = f(Z_{mis}|Z_{obs})$. Therefore, with the ignorability assumption, we do not need to model the distribution of R in multiple imputations. In other words, in the imputation step when we intend to generate M imputed data sets for Z_{mis}, we can draw from $f(Z_{mis}|Z_{obs})$ that drops out the modeling of Z.

However, even drawing from $f(Z_{mis}|Z_{obs})$ can be difficult in some cases such as when the data have general missing-data patterns, when the joint likelihood function is complex, and when the predictive distribution is intractable. Data augmentation (Tanner and Wong, 1987), a method built on Markov Chain Monte Carlo (MCMC), is a natural solution to this issue and is widely used in MI. Bayesian statistics treats Z_{mis} as additional parameters to θ. The goal is to sample from the joint distribution for Z_{mis} and θ given Z_{obs}. The iterative procedure begins with an initial value of Z_{mis}, followed by drawing θ from $f(\theta|Z_{obs}, Z_{mis})$. At the t-th iteration, given the current value of the parameter $\theta^{(t)}$, we draw new values for Z_{mis}, say $Z_{mis}^{(t+1)}$, from the conditional predictive distribution $f(Z_{mis}|Z_{obs}, \theta^{(t)})$. We then draw a new value for θ from the complete-data posterior distribution $f(\theta|Z_{obs}, Z_{mis}^{(t+1)})$. As the iterative chain reaches a stationary state, Li (1988) show that the distribution of $\theta^{(t)}$ will approximate $f(\theta|Z_{obs})$ and the distribution of $Z_{mis}^{(t)}$ will approximate $f(Z_{mis}|Z_{obs})$. In addition to data augmentation strategy, there are some other ways of implementing imputations and one of the frequently used approach is hot-deck, as we introduce before. Most hot-deck procedures match the non-respondents with similar responding subjects and replace the missing values of the former with the observed values of the latter. However, hot-deck methods are not proper imputation methods as defined in Rubin (1987). A proper imputation method should incorporate appropriate variability among the M sets of complete data sets and yield consistent results.

When achieving M complete data sets, $(Z_{obs}$,$Z_{mis}^{(1)})$,...,$(Z_{obs}, Z_{mis}^{(M)})$, the next step is to analyze these data sets using common complete-data methods in order to get parameters estimates and standard errors. Such an analysis is conducted using one complete data set at a time, aiming to estimate the parameter of interest such as regression coefficients. Of note, an analyst should be able to distinguish the parameter that the individual desires to estimate and the parameter in Equation 1.1, as the former is from the analysis model and the latter is from the imputation model. Meng (1994) argues that the imputation model should be more general than the analysis model and be relatively "rich" in order to be congenial with lots of different analysis models of interest. Let Q be the parameter of interest and under the setting of congeniality, Q should be some function of θ. We denote the estimate of Q from the analysis using $(Z_{obs}, Z_{mis}^{(m)})$ by $\widehat{Q}^{(m)}$, and the corresponding variance estimates by $\widehat{V}^{(m)}$, where $m = 1, ..., M$.

After we gather the results (i.e., estimates and standard errors) from the analysis step, we can combine them and get a final result using Rubin's rules (Rubin, 1987). The combined estimate of Q is

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \widehat{Q}^{(m)}$$

and its variance estimate is

$$T = (1 + \frac{1}{M})B + \bar{V}$$

, where the between-imputation variance $B = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{Q}^{(m)} - \bar{Q})^2$ and the within-imputation variance $\bar{V} = \frac{1}{M} \sum_{m=1}^{M} \widehat{V}^{(m)}$. When the missing-data mechanism is ignorable and the imputation method is proper, MI will lead to consistent and asymptotically efficient estimates. These advantages make MI one of the leading methods in handling missing data.

1.2.3 Inverse Probability Weighting

Most standard complete-data analyses treat all subjects as equally important. However, in some situations such as meta-analysis and survey sampling, it may be proper to vary the weights given to different subjects. Horvitz and Thompson (1952) first bring the idea of weighting into the missing data research and named it inverse probability weighting (IPW). They argue that the bias of only using the complete cases can be corrected, in some circumstances, when weighting each respondent by the inverse of the probability of being a respondent. Consider a generalized linear model $E(Y) = g^{-1}(X, \theta)$ with missing covariate X, IPW method tries to solve the weighted estimating equation $U_{IPW}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i w_i U_i(\theta)$, where $U_i(\theta)$ is the first derivative of the log likelihood function with respect to θ, and R_i is a binary observe indicator and $w_i = 1/Pr(R_i = 1)$ is the weight for subject i. Note that if $w_i = 1$ for all i, IPW estimator becomes the estimator from the CC analysis. Usually, the weights are unknown and we can use a parametric function to model them, for example, $\text{logit}(\frac{1}{w_i}) = \beta_0 + \beta_1 Y + \beta_2 X_{obs}$, where β's are unknown parameters. We may obtain the estimated $\hat{w}_i$ by plugging in the maximum likelihood estimator of β. Wang et al. (1997) apply nonparametric method (kernel smoother) to estimate the weights. They prove that the estimators using known weights, estimated weights from the parametric model or the nonparametric model are all consistent, while the one using known weights is less efficient than others. The asymptotic variances of IPW estimators are provided in their paper, however, people usually use bootstrap methods to approximate the variances for practical purposes.

Comparing to MI, IPW specifies a missingness model for the probability that a subject is a complete case. However, such model may lead to unstable weights and therefore inefficiency of the estimators. One approach to fixing it is by weight stabilization. The weight is stabilized by being replaced with a fraction with the numerator and denominator corresponding to two estimated weights from different models. Another way to improve efficiency of IPW is through augmented IPW (AIPW) (Robins et al., 1994; Scharfstein et al., 1999). AIPW is a hybrid of IPW and imputation, that possesses the property of doubly robust. In other words, AIPW is consistent either the weight model or imputation model is correctly specified. Moreover, it is fully efficient if both models are correct. However, in the presence of general missing-data patterns and complex data, IPW methods lose their power since modeling the weights is difficult. Seaman and White (2013) give a nice review of IPW and compare it with MI.

1.3 Existing Methods for Handling Incomplete Big Data

Standard statistical methods for analyzing incomplete data, including the CC analysis, MI and IPW, gain their popularity on low dimensional data. However, in the presence of incomplete big data, their application is problematic especially when the data are complex and high-dimensional. In practice, it is very common to have large number of variables with missing values. In such cases, standard MI methods usually face severe challenges because the imputation models for missing values are complicated and the intensive computing is most likely insurmountable. In the past two decades, with the increasing collection of data, modern techniques for handling incomplete big data have drawn great attention and both nonparametric and parametric methods have been developed.

1.3.1 Nonparametric Methods for Incomplete Big Data

Troyanskaya et al. (2001) propose a k-nearest neighbor single-imputation method (KNNimpute) and utilize subjects in the neighborhood to impute missing values, by a weighted average of observed values. The selection of neighbors is based on the Euclidean distance. Liao et al. (2014) extend the KNNimpute to account for the information contained in the nearest variables. Specifically, they proposed four variations of KNNimpute among which, two approaches are shown to outperform a standard MI method in the simulation study. However, an inevitable problem of incorporating standard k-nearest neighbor into imputation methods is the specification of k as their performance highly depends on k. Tutz and Ramzan (2015) propose a new method that can automatically select the relevant neighbors, instead of depending on k. Such a method based on a distance which uses the correlation among variables preforms well, especially when there are a large number of variables.

Random forest, an ensemble method for classification and regression tasks, is widely used by researchers to cope with big data. As a nonparametric method, it allows complex interactive and non-linear relations between variables of different types. Stekhoven and Bühlmann (2012) propose a random-forest-based imputation method (missForest) in which, for each variable with missing values, a random forest model is fitted using the observed values of that variable as the response and the remaining variables as predictors. By inheriting the power of random forest, missForest exhibits attractive computational efficiency and performs particularly well, compared with other parametric methods in the presence of mixed-type big data. Although the author argued that missForest intrinsically constitutes multiple imputation since many unpruned decision trees are averaged in the random forest procedure, it is not a proper imputation by the definition of Rubin (1987). This is because missForest predicts each missing value using a exact value rather than the random draw from a distribution. In a more recent paper, Shah et al. (2014) investigate the improper issue of missForest and proposed a new method (MICE-RF) within the multiple imputation by chained equation framework, where multiple bootstrap samples of the original data are used to generate multiple imputed data sets. The step of bootstrapping accommodates the sampling variation and thus ensures proper imputations. Valdiviezo and Van Aelst (2015) compare several strategies such as single and multiple imputations to handle missing data when using tree-based prediction methods, with a focus on their practical applications. Such prediction methods include the classification and regression tree proposed by Breiman et al. (1984), the conditional inference forests proposed by Hothorn et al. (2006), random forests, conditional inference forests developed by Hothorn et al. (2006), and etc. The paper recommended that if the data have small amounts of missing values, the CC analysis or the single imputation method is sufficient; in the presence of large incomplete data, MI method using the conditional inference forests as the prediction model is suggested.

Other novel nonparametric methods include a multiple imputation approach replying on the combination of a multiplayer perceptron and k-nearest neighbors (Silva-Ramírez et al., 2015). However, this method can only deal with data of monotone missing-data patterns.

1.3.2 Parametric Methods for Incomplete Big Data

Städler and Bühlmann (2012) propose a likelihood approach to solve the so-called matrix completion problem, where the goal is to recover a matrix from an incomplete set of entries. They assume a multivariate normal model with p-dimensional covariance matrix and presented an EM algorithm that maximizes the l_1-penalized observed log-likelihood. The method (MissGLasso) is found to be always better than KNNimpute in their simulation study. However, the E-step in the EM algorithm is rather complex and MissGLasso strongly depends on the MAR assumption and cannot handle data that are MNAR. In 2014, the authors develop a more efficient algorithm called MissPALasso (Städler et al., 2014), focusing on improving MissGLasso in two aspects: inefficiency of the EM algorithm and non-sparseness of the regression coefficients.

Song and Belin (2004) provide a procedure for multiple imputation based on a common factor model to reduce the dimension of the parameters in a multivariate normal model. The assumed factor model has k underlying factors and can be described as $Z_i = \alpha + \boldsymbol{W}_i\beta + \epsilon_i$ for $i = 1, 2, ..., n$, where α is a $1 \times p$ mean vector, $\boldsymbol{W}_i$ is a $1 \times k$ factor score vector, β is a $k \times p$ factor loading matrix, and ϵ_i follows a multivariate normal distribution. With the complete-data likelihood, the parameters and factor scores $\boldsymbol{W}$, as well as missing items can be simulated using a Gibbs sampler, which is viewed as an application of data augmentation.

In the presence of mixed continuous and binary data, Audigier et al. (2016) propose a new single imputation method based on a principal component method dedicated to mixed data: the factorial analysis for mixed data (FAMD). FAMD can reduce the dimensionality of the data by providing a subspace that best represents the data, and account for the influence of the continuous and the categorial variables in the analysis as well. However, as a single imputation method, the proposed method does not account for the uncertainty brought by the imputation. Another parametric imputation method for high-dimensional mixed data is presented by He and Belin (2014). It is a joint modeling approach, where latent variables are used to model the binary variables. They use the generalized parameter-expanded Metropolis-Hastings algorithm (Boscardin et al., 2008) to sample the mixed covariance-correlation matrix for the joint distribution of continuous and latent variables associated with binary variables. The simulation study indicates that the multiple imputation method can adapt to different covariance structures when the data are MAR.

1.4 Distributed Analysis of Big Data

To improve efficiency, it would be beneficial for different organizations with the same research topic or sub-units within an organization to collaborate and share individual views and further outcomes on the topic. Such a collaboration would contribute a larger sample size and a more representative population. For instance, a healthcare system consists of various public and private data collection systems such as health surveys, administrative enrollment and billing records. These collection systems are usually conducted by different entities, including hospitals, physicians, and insurance companies. Very few entities are capable by themselves to gather all characteristics for the entire population of patients. Thus, such a circumstance motivates the need of statisticians to develop essential techniques to analyze data from multiple sources. One would argue that we can simply pool the data from multiple sources and perform standard statistical analyses using the pooled data, which is definitely the best way of using all the information from the data. However, directly merging the data may not be allowed in practice due to the following concerns. First, policies and regulations do not authorize data to be shared across organizations. For example, Veteran's Health Administration's policy restricts their data to be only at internal facilities. Second, privacy of sharing sensitive data such as hospital medical records and health insurance bill remains to be an essential issue because directly merged data sets may result in the disclosure of susceptible information by malicious parties. As a result, we introduce distributed analyses that achieve the desired statistical goal without merging exact data from multiple sources, which preserves the confidentiality of the data sets. The distributed analyses allow some computations to be conducted locally, which largely improve the computation efficiency. In Chapter 3 and 4, our proposed methods for handling distributed incomplete data depend on the distributed analyses illustrated below. To begin with, distributed analyses are typically performed on two forms of data: horizontally partitioned data and vertically partitioned data, as illustrated in Figure 1.3.

1.4.1 Distributed Analysis for Horizontally Partitioned Data

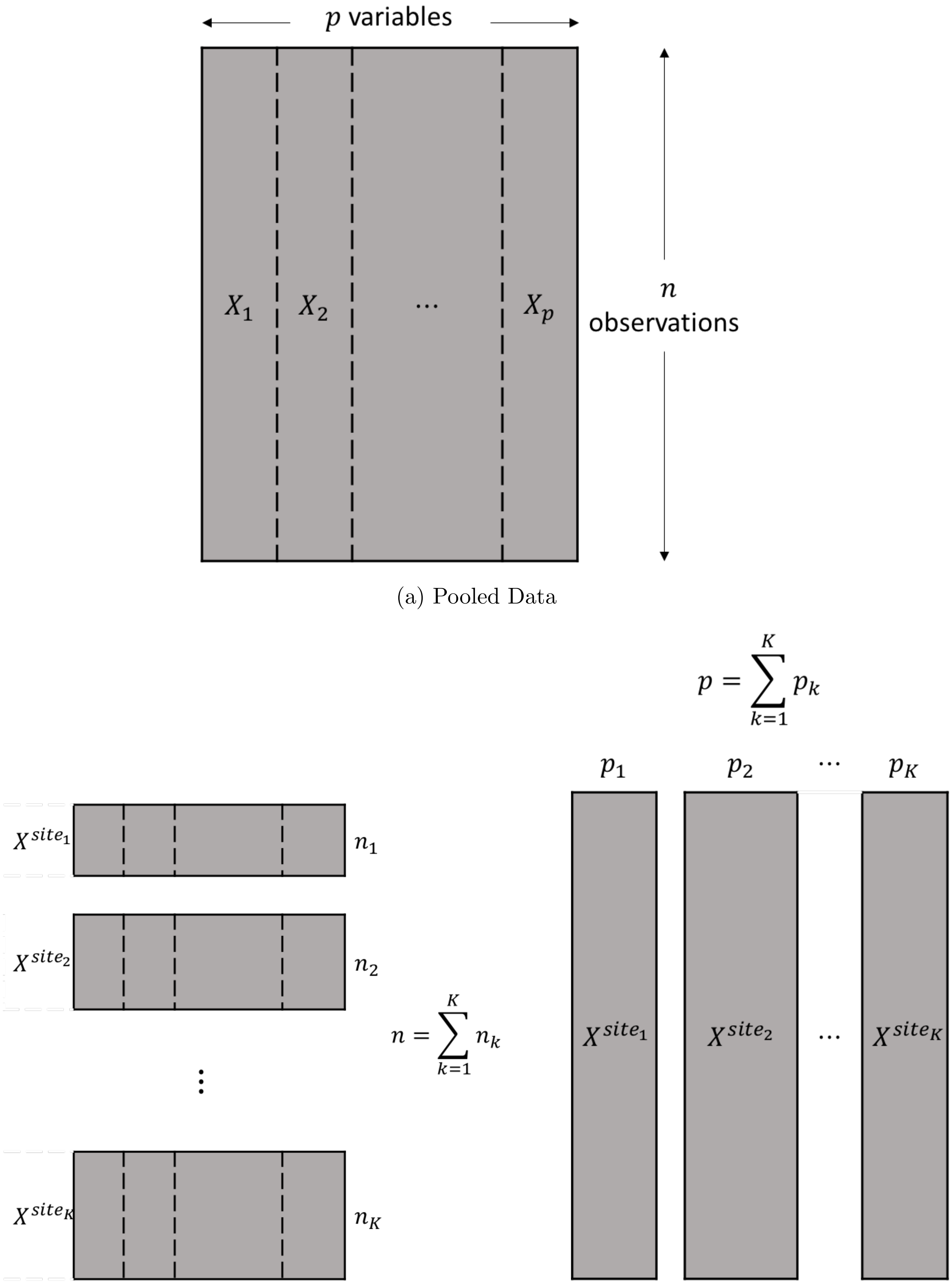
Horizontally partitioned data emerge when several institutions (sites) gather the same collection of information on various entities. For example, the Behavioral Risk Factor Surveillance System (BRFSS) operates state-wise telephone surveys to gather the same set of information on health-related risk behaviors across the United States. Correspondingly, the BRFSS data within each state are considered horizontally partitioned.

Assume the data consist response vector Y and design matrix X, such that:

$$Y = \begin{pmatrix} Y^{site_1} \\ \vdots \\ Y^{site_K} \end{pmatrix}, \quad X = \begin{pmatrix} X^{site_1} \\ \vdots \\ X^{site_K} \end{pmatrix},$$

where Y^{site_k} and X^{site_k} $(1 \leq k \leq K)$ are data from institution k.

We first consider distributed linear regression for horizontally partitioned data. Sup-



(a) Pooled Data

(b) Horizontally Partitioned Data

(c) Vertically Partitioned Data

Figure 1.3: Pooled data and two types of distributed data

pose that Y follows a normal distribution:

$$Y = X\beta + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. The estimate of the parameter is

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.2}$$

In our distributed environment, the pooled data Y and X are not accessible due to certain concerns. Nonetheless, two products $X^T X$ and $X^T Y$ from Equation 1.2 can be obtained from local sites as follows,

$$X^T X = \sum_{k=1}^{K} (X^{site_k})^T X^{site_k}, \quad X^T Y = \sum_{k=1}^{K} (X^{site_k})^T Y^{site_k}$$

In this secure summation process, only the summary statistics $((X^{site_k})^T X^{site_k}$, $(X^{site_k})^T Y^{site_k})$ are calculated in parallel locally and then transmitted to a mater site. Figure 1.4 shows how the summary statistics transmit to the master site in a linear regression with 3 worker sites.

We now turn to a distributed logistic regression for horizontally partitioned data. Assume Y is a binary variable:

$$\mathrm{logit}(Pr(Y = 1)) = X\beta.$$

Since the estimate of β cannot be found in a closed form, a commonly used way is called Newton-Raphson that approximates $\hat{\beta}$ iteratively in the following way:

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (X^T W^{(t)} X)^{-1} X^T (Y - \mu^{(t)}), \tag{1.3}$$

where $W^{(t)} = \mathrm{diag}(\pi_i^{(t)}(1 - \pi_i^{(t)}))$, $\mu^{(t)} = (\pi_i^{(t)})$ and $\pi_i^{(t)}$ is the probability of $Y = 1$ for the

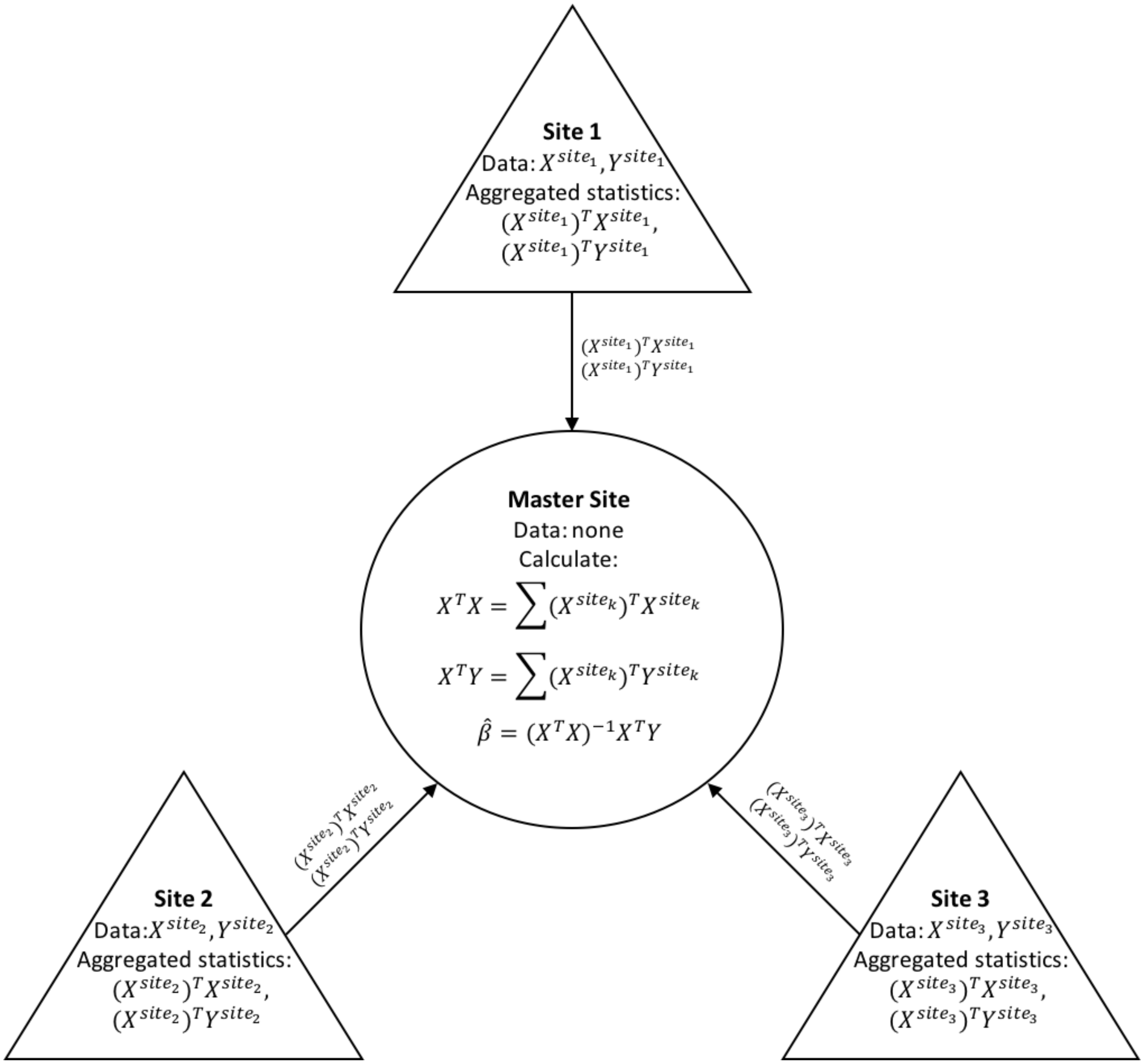

Site 1
Data: X^{site_1}, Y^{site_1}
Aggregated statistics:
$(X^{site_1})^T X^{site_1}$,
$(X^{site_1})^T Y^{site_1}$
$(X^{site_1})^T X^{site_1}$
$(X^{site_1})^T Y^{site_1}$
Master Site
Data: none
Calculate:
$X^T X = \sum (X^{site_k})^T X^{site_k}$
$X^T Y = \sum (X^{site_k})^T Y^{site_k}$
$\hat{\beta} = (X^T X)^{-1} X^T Y$
$(X^{site_2})^T X^{site_2}$
$(X^{site_2})^T Y^{site_2}$
$(X^{site_3})^T X^{site_3}$
$(X^{site_3})^T Y^{site_3}$
Site 2
Data: X^{site_2}, Y^{site_2}
Aggregated statistics:
$(X^{site_2})^T X^{site_2}$,
$(X^{site_2})^T Y^{site_2}$
Site 3
Data: X^{site_3}, Y^{site_3}
Aggregated statistics:
$(X^{site_3})^T X^{site_3}$,
$(X^{site_3})^T Y^{site_3}$

Figure 1.4: Distributed linear regression on data that are horizontally distributed across $K = 3$ sites

i^{th} observation in iteration t. We can show that in the presence of horizontally partitioned data, the update step 1.3 can be easily performed using local summary statistics:

$$X^T W^{(t)} X = \sum_{k=1}^{K} (X^{site_k})^T W^{(t)site_k} X^{site_k} \tag{1.4}$$

$$X^T (Y - \mu^{(t)}) = \sum_{k=1}^{K} (X^{site_k})^T (Y^{site_k} - \mu^{(t)site_k}) \tag{1.5}$$

where $\mu^{(t)site_k}$ is the vector of probabilities of $Y = 1$ for subjects in institution k in iteration t, and $W^{(t)site_k}$ is the diagonal matrix of $\pi_i^{(t)}(1-\pi_i^{(t)})$ for all subject i belongs to institution k. Figure 1.5 illustrates the quantities transmission between master and local sites in a distributed logistic regression assuming $K = 3$ sites participate in the collaboration. Of note, the distributed linear regression only requires an one-time transmission of summary statistics to the master site, while the distributed logistic regression involves the iterative two-way transmission of quantities between master and local sites.

Wu et al. (2012) systematically investigate the above distributed logistic regression and named it Grid Binary LOgistic REgression (GLORE). In addition to parameter estimates, the GLORE model integrates decomposable partial elements or non-privacy sensitive prediction values to obtain the variance covariance matrix, the goodness-of-fit test statistic, and the area under the receiver operating characteristic (ROC) curve. Jiang, Li, Wang, Wu, Xue, Ohno-Machado and Jiang (2013) implement the distributed logistic regression using JAVA and provide the corresponding easy-to-use web service for researchers.

1.4.2 Distributed Analysis for Vertically Partitioned Data

When data are vertically partitioned, to simplify our problem, we assume the data are from two sites/institutions. Suppose institution 1 has a data set A for n subjects and institution 2 has another data set B for the same subjects, thus $X = (A, B)$. We further make an assumption that both institutions know the outcome variable Y. Considering a

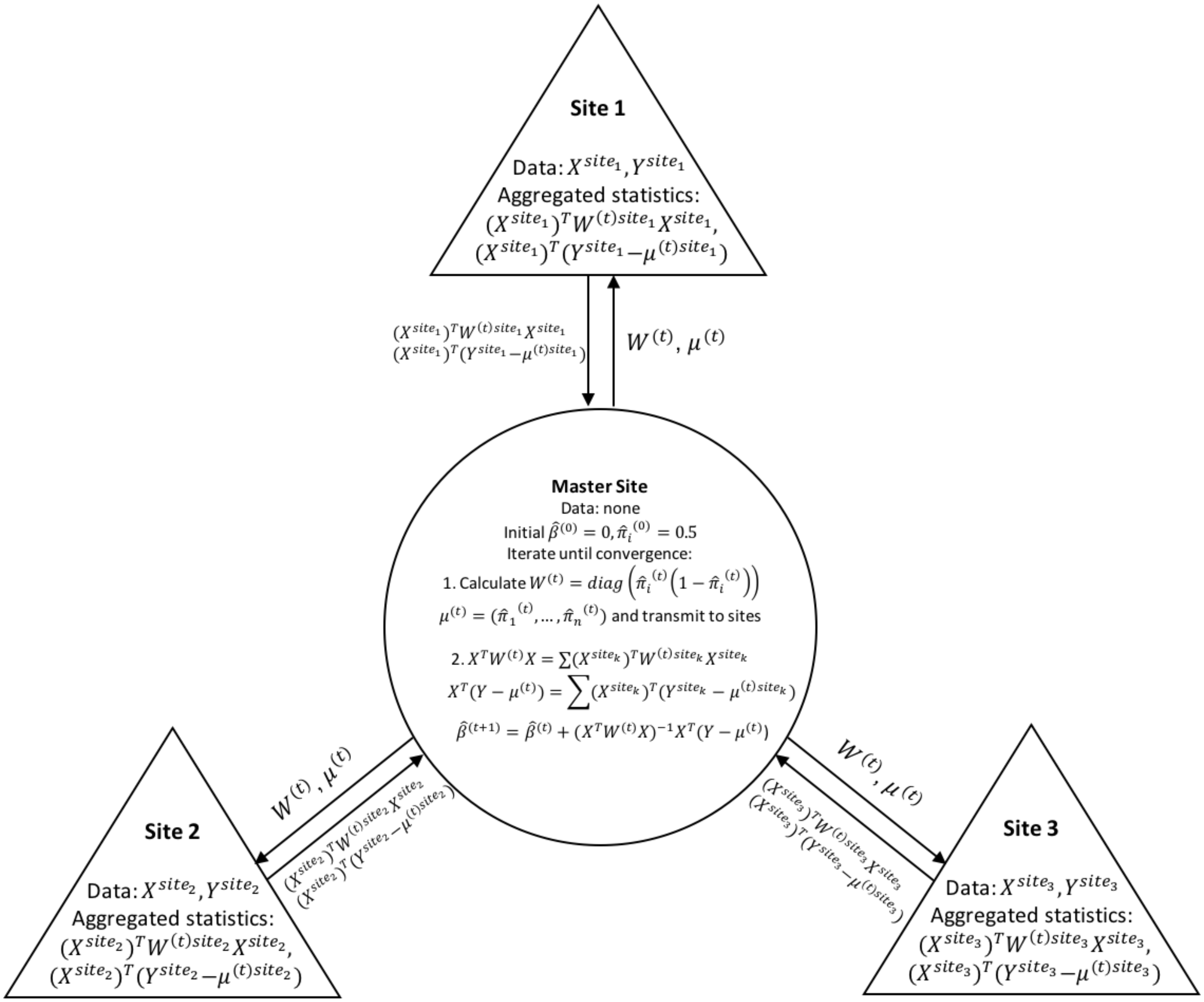

Site 1
Data: X^{site_1}, Y^{site_1}
Aggregated statistics:
$(X^{site_1})^T W^{(t)site_1} X^{site_1}$,
$(X^{site_1})^T (Y^{site_1} - \mu^{(t)site_1})$
$(X^{site_1})^T W^{(t)site_1} X^{site_1}$
$(X^{site_1})^T (Y^{site_1} - \mu^{(t)site_1})$
$W^{(t)}, \mu^{(t)}$
Master Site
Data: none
Initial $\hat{\beta}^{(0)} = 0, \hat{\pi}_i^{(0)} = 0.5$
Iterate until convergence:
1. Calculate $W^{(t)} = diag\left(\hat{\pi}_i^{(t)}\left(1 - \hat{\pi}_i^{(t)}\right)\right)$
$\mu^{(t)} = (\hat{\pi}_1^{(t)}, \dots, \hat{\pi}_n^{(t)})$ and transmit to sites
2. $X^T W^{(t)} X = \sum (X^{site_k})^T W^{(t)site_k} X^{site_k}$
$X^T (Y - \mu^{(t)}) = \sum (X^{site_k})^T (Y^{site_k} - \mu^{(t)site_k})$
$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (X^T W^{(t)} X)^{-1} X^T (Y - \mu^{(t)})$
$W^{(t)}, \mu^{(t)}$
$(X^{site_2})^T W^{(t)site_2} X^{site_2}$
$(X^{site_2})^T (Y^{site_2} - \mu^{(t)site_2})$
Site 2
Data: X^{site_2}, Y^{site_2}
Aggregated statistics:
$(X^{site_2})^T W^{(t)site_2} X^{site_2}$,
$(X^{site_2})^T (Y^{site_2} - \mu^{(t)site_2})$
$W^{(t)}, \mu^{(t)}$
$(X^{site_3})^T W^{(t)site_3} X^{site_3}$
$(X^{site_3})^T (Y^{site_3} - \mu^{(t)site_3})$
Site 3
Data: X^{site_3}, Y^{site_3}
Aggregated statistics:
$(X^{site_3})^T W^{(t)site_3} X^{site_3}$,
$(X^{site_3})^T (Y^{site_3} - \mu^{(t)site_3})$

Figure 1.5: Distributed logistic regression on data that are horizontally distributed across $K = 3$ sites

linear regression on Y with X, the goal is to obtain $\hat{\beta}$ of Equation 1.2 when the data are vertically partitioned. Note that,

$$X^T X = \begin{pmatrix} A^T A & A^T B \\ B^T A & B^T B \end{pmatrix}$$

, and

$$X^T Y = \begin{pmatrix} A^T Y \\ B^T Y \end{pmatrix}$$

Du et al. (2004) develop secure multi-party computation protocols for privacy-preserving calculations of matrix product and matrix inverse. By using their secure technique, we can obtain $\hat{\beta}$ without passing A to institute 1 and passing B to institute 2.

An alternative approach of distributed linear regression is proposed by Sanil et al. (2004). Details will be discussed in Chapter 4.

As for logistic regression, Slavkovic et al. (2007) propose an algorithm to aggregate information among institutions through secure multi-party computation protocols. However, the algorithm induces very high computational cost and is not scalable as K is large. Li et al. (2015) publish a distributed algorithm to solve the maximum likelihood problem by dual optimization. Their method is shown to be more efficient from the simulation study.

1.4.3 Missing Data in Distributed Analysis

The above-mentioned techniques of distributed analysis assume that the data are complete. However, it is especially common that data from multiple sources are subject to missing values. Based on our knowledge, Jagannathan and Wright (2008) is the first and only paper that investigated missing data in a distributed analysis. In that paper, the author propose a privacy-preserving single imputation algorithm based on decision trees.

The method can deal with the missing data problem when the data are collected from two sources and observed to have a univariate missing-data pattern.

Chapter 2

Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data

2.1 Introduction

Missing data are often encountered for various reasons in biomedical research and present challenges for data analysis. It is well known that inadequate handling of missing data may lead to biased estimation and inference. A number of statistical methods have been developed for handling missing data. Largely due to its ease of use, multiple imputation (MI)(Rubin, 1987; Little and Rubin, 2002) has been arguably the most popular method for handling missing data in practice. The basic idea underlying MI is to replace each missing data point with a set of values generated from its predictive distribution given observed data and to generate multiply imputed datasets to account for uncertainty of imputation. Each imputed data set is then analyzed separately using standard complete-data analysis methods and the results are combined across all imputed data sets using Rubin's rules (Rubin, 1987; Little and Rubin, 2002). MI can be readily conducted using available software packages van Buuren and Groothuis-Oudshoorn (2011); Raghunathan et al. (2001); Su, Gelman, Hill and Yajima (2011) in a wide range of situations and has been investigated extensively in many settings Harel and Zhou (2007); He et al. (2011); Hsu et al. (2004); Little and An (2004); Long et al. (2012); Qi et al. (2010); Zhang and Little (2008). Most of the existing MI methods rely on the assumption of *missingness at random* (MAR) (Little and Rubin, 2002), i.e., missingness only depends on observed data; our current work also focuses on MAR. In recent years, the amount of data has increased considerably in many applications such as omic data and electronic health record data. In particular, the high dimensions in omic data may cause serious problems to MI in terms of applicability and accuracy. In what follows, we first describe some challenges of MI in the presence of high-dimensional data and explain why regularized regressions are suitable in this setting, and then review existing MI methods for general missing data patterns and propose their extensions for high-dimensional data.

Advances in technologies have led to collection of high-dimensional data such as omics data in many biomedical studies where the number of variables is very large and missing data are often present. Such high-dimensional data present unique challenges to MI. When conducting MI, Meng (1994) suggests imputation models be as general as data allow them to be, in order to accommodate a wide range of statistical analyses that may be conducted using multiply imputed data sets. However, in the presence of high-dimensional data, it is often infeasible to include all variables in an imputation model. As such, machine learning and model trimming techniques have been used in building imputation models in these settings. Stekhoven and Bühlmann (2012) propose a random forest-based algorithm for missing data imputation called missForest. Random forest utilizes bootstrap aggregation of multiple regression trees to reduce the risk of overfitting, and combines the predictions from trees to improve accuracy of predictions (Breiman, 2001). Shah et al. (2014) suggest a variant of missForest and compared it to parametric imputation methods. They showed that their proposed random forest imputation method was more efficient and produced narrower confidence intervals than standard MI methods. Liao et al. (2014) develop four variations of K-nearest-neighbor (KNN) imputation methods. However, these methods are improper in the sense of Rubin (1987) since they do not adequately account for the uncertainty of estimating parameters in the imputation models. Improper imputation may lead to biased parameter estimates and inference in subsequent analyses. In addition, KNN methods are known to suffer from the curse of dimensionality (Marimont and Shapiro, 1979; Stone, 1980) and hence may not be suitable for high-dimensional data. Apart from random forest and KNN, regularized regression, which allows for simultaneous parameter estimation and variable selection, presents another option for building imputation models in the presence of high-dimensional data. The basic idea of regularized regression is to minimize the loss function of a regression, subject to some penalties. Different penalty specifications give rise to various regularized regression methods. Zhao and Long (2013) investigate the use of regularized regression for MI including lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) (EN), and adaptive lasso (Zou, 2006) (Alasso). They also develop MI using a Bayesian lasso approach. However, they focus on the setting where only one variable has missing values. There has been limited work on MI methods for general missing data patterns where multiple variables have missing values in the presence of high-dimensional data.

To handle general missing data patterns, there are two MI approaches, namely, MI based on joint modeling (JM) (Schafer, 1997) and MI based on fully conditional specifications, also known as multiple imputation by chained equations (MICE), which has been implemented independently by van Buuren et al. (1999) and Raghunathan and Siscovick (1996). While JM has strong theoretical justifications and works reasonably well for low-dimensional data, its performance deteriorates as the data dimension increases (Van Buuren, 2007) and it is difficult to extend to high-dimensional data. MICE involves specifying a set of univariate imputation models. Since each imputation model is specified for one partially observed variable conditional on the other variables, it simplifies the modeling process. While MICE lacks theoretical justifications except for some special cases (Liu et al., 2013; Zhu and Raghunathan, 2014), it has been shown to achieve satisfactory performance in extensive numerical studies and empirical examples. White et al. (2011) provides a nice review and guidance for MICE. It is worth mentioning that standard MICE methods cannot handle high-dimensional data. For example, the MICE algorithms implemented by Buuren and Groothuis-Oudshoorn (2011); Su, Gelman, Hill, Yajima et al. (2011) cannot handle the prostate cancer data used in our data analysis and the high-dimensional data generated in our simulations, as shown in later sections. As such, we focus on extending MICE to high-dimensional data settings for handling general missing data patterns.

2.2 Methodology

Suppose that our data set Z has p variables, $Z_1, ..., Z_p$. Without loss of generality, we assume that the first l ($l \leq p$) variables contain missing values. Suppose the data consist of n observations and we have r_j observed values in variable Z_j. We denote the observed components and missing components for variable j by $Z_{j,obs}$ and $Z_{j,mis}$. Let $Z_{-j} = (Z_1, ..., Z_{j-1}, Z_{j+1}, ..., Z_p)$ be the collection of the $p-1$ variables in Z except Z_j. Let $Z_{-j,obs}$ and $Z_{-j,mis}$ denote the two components of Z_{-j} corresponding to the complement data of $Z_{j,obs}$ and $Z_{j,mis}$.

2.2.1 Multiple imputation by chained equations

Let the hypothetically complete data Z be a partially observed random draw from a multivariate distribution $f(Z|\theta)$. We assume that the multivariate distribution of Z is completely specified by the unknown parameters θ. The standard MICE algorithm obtains a posterior distribution of θ by sampling iteratively from conditional distributions of the form $f(Z_1|Z_{-1}, \theta_1), ..., f(Z_l|Z_{-l}, \theta_l)$. Note that the parameters $\theta_1, ..., \theta_l$ are specific to the conditional densities, which might not determine the unique 'true' joint distribution $f(Z|\theta)$.

To be specific, MICE starts with a simple imputation, such as imputing the mean, for every missing value in the data set. Initial values are denoted by $Z_1^{(0)}, ..., Z_l^{(0)}$. Then given values $Z_1^{(m-1)}, ..., Z_l^{(m-1)}$ at iteration $m-1$, for variable j, new parameter estimates $\hat{\theta}_j^{(m)}$ of the next iteration are generated from

$$f(\theta_j | Z_{j,obs}, Z_1^{(m)}, ..., Z_{j-1}^{(m)}, Z_{j+1}^{(m-1)} ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p)$$

through a regression model. Then the missing values $Z_{j,mis}$ for Z_j are replaced with predicted values from the regression model with model parameter $\hat{\theta}_j^{(m)}$. Note that when Z_j is subsequently used as a predictor in the regression model for other variables that have missing values, both the observed and predicted values are used. These steps are repeated for each variable with missing values, that is, Z_1 to Z_l. We call the cycling imputing through Z_1 to Z_l one iteration. At the end of each iteration, all missing values are replaced by the predictions from regression models that expose the relationships observed in the data. We then repeat the procedures iteratively until convergence. The complete algorithm can be described as follows:

$$
\begin{aligned}
\widehat{\theta}_1^{(m)} &\sim f(\theta_1|Z_{1,obs}, Z_2^{(m-1)}, ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p) \\
Z_{1,mis}^{(m)} &\sim f(Z_{1,mis}|Z_2^{(m-1)}, ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p, \widehat{\theta}_1^{(m)}) \\
&\vdots \\
\widehat{\theta}_j^{(m)} &\sim f(\theta_j|Z_{j,obs}, Z_1^{(m)}, ..., Z_{j-1}^{(m)}, Z_{j+1}^{(m-1)} ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p) \qquad (2.1a) \\
Z_{j,mis}^{(m)} &\sim f(Z_{j,mis}|Z_1^{(m)}, ..., Z_{j-1}^{(m)}, Z_{j+1}^{(m-1)}, ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p, \widehat{\theta}_j^{(m)}) \qquad (2.1b) \\
&\vdots
\end{aligned}
$$

Note that while the observed data Z_{obs} do not change in the iterative updating procedure, the missing data Z_{mis} do change from one iteration to another. After convergence, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.

In the case of high-dimensional data, where $p > r_j$ or $p \approx r_j$, it is not feasible to fit the imputation model (2.1a) using traditional regressions. In the following two subsections, we provide details of two approaches to apply regularized regression techniques in the presence of high-dimensional data for general missing data patterns.

2.2.2 Direct use of regularized regression for multiple imputation

For variable Z_j, our goal is to fit the imputation model (2.1a) using r_j cases with observed Z_j. Assume q variables in $Z_{-j,obs}$ are associated with $Z_{j,obs}$ and we denote the set of them by $\mathcal{S}$, which we call the true active set. We define the subset of predictors that are selected to impute Z_j as the active set by $\widehat{\mathcal{S}}$, and denote the corresponding design matrix as

$Z_{\widehat{\mathcal{S}},obs}$.We first consider an approach where a regularization method is used to conduct both model trimming and parameter estimation and a bootstrap step is incorporated to simulate random draws from $f(\theta \mid Z_{j,obs}, Z_{\widehat{\mathcal{S}},obs})$. This approach is referred to as MICE through the direct use of regularized regression (MICE-DURR). The purpose of the boostrap is to accommodate sampling variation in estimating population regression parameters, which is part of ensuring that imputations are proper (Shah et al., 2014). In the m-th iteration and for variable j,$(j = 1, ..., l)$, define $\mathbf{W}_j^{(m)} = \{Z_1^{(m)}, ..., Z_{j-1}^{(m)}, Z_{j+1}^{(m-1)}, ..., Z_l^{(m-1)}, Z_{l+1}, ..., Z_p\}$. Denote by $\mathbf{W}_{j,mis}^{(m)}$ the component of $\mathbf{W}_j^{(m)}$ corresponding to $Z_{j,mis}$. The algorithm can be described as follows:

(1) Generate a bootstrap data set $\{\mathbf{W}_j^{*(m)}, Z_j^*\}$ of size n by randomly drawing n observations from $\{\mathbf{W}_j^{(m)}, Z_j^{(m-1)}\}$ with replacement. Denote the observed values of Z_j^* by $Z_{j,obs}^*$ and the corresponding component of $\mathbf{W}_j^{*(m)}$ by $\mathbf{W}_{j,obs}^{*(m)}$.

(2) Regarding $Z_{j,obs}^*$ as the outcome and $\mathbf{W}_{j,obs}^{*(m)}$ as predictors, use a regularized regression method to fit the model and obtain $\widehat{\theta}_j^{(m)}$. Note that $\widehat{\theta}_j^{(m)}$ is considered a random draw from $f(\theta_j | Z_{j,obs}, Z_{-j,obs})$.

(3) Predict $Z_{j,mis}$ with $Z_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $f(Z_{j,mis} | \mathbf{W}_{j,mis}^{(m)}, \widehat{\theta}_j^{(m)})$, noting that imputation is conducted on the original data set $\mathbf{W}_{j,mis}^{(m)}$, not the bootstrap data set $\mathbf{W}_j^{*(m)}$.

We conduct the above procedure for l variables that have missing values in one iteration and repeat iteratively to obtain M imputed data sets. Subsequently, standard complete-data analysis can be applied to each one of the M imputed data sets.

We make our approach clear by linking the above three steps to the MICE algorithm. In the first step, we bootstrap the data from the last iteration to ensure that the following imputations are proper. In the second step, we use regularized regressions to fit model (2.1a) and obtain an estimate of θ_j. Then, we use this estimate to predict the missing values from the model (2.1b). Details of MICE-DURR for three types of data can be found as Supplementary Method S1 online.

2.2.3 Indirect use of regularized regression for multiple imputation

MICE-DURR uses regularized regression for both model trimming and parameter estimation. An alternative approach to MICE-DURR is to use a regularization method for model trimming only and then followed by a standard multiple imputation procedure using the estimated active set ($\widehat{\mathcal{S}}$), say, through a maximum likelihood inference procedure. We refer to this approach as MICE through the indirect use of regularized regression (MICE-IURR). Suppose $\mathbf{W}_j^{(m)}$ is defined as above. Denote by $\mathbf{W}_{j,obs}^{(m)}$ the component of $\mathbf{W}_j^{(m)}$ corresponding to $Z_{j,obs}$. At the m-th iteration and for variable Z_j, the algorithm of the MICE-IURR approach is as follows:

(1) We use a regularized regression method to fit a multiple linear regression model regarding $Z_{j,obs}$ as the outcome variable and $\mathbf{W}_{j,obs}^{(m)}$ as the predictor variable, and identify the active set, $\widehat{\mathcal{S}}_j^{(m)}$. Let $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ denote the subset of $\mathbf{W}_j^{(m)}$ that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}$.

(2) Approximate the distribution of $f(\theta_j \mid Z_{j,obs}, \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs})$ by using a standard inference procedure such as maximum likelihood.

(3) Predict $Z_{j,mis}$: randomly draw $\widehat{\theta_j}^{(m)}$ from $f(\theta_j \mid Z_{j,obs}, \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs})$ and subsequently predict $Z_{j,mis}$ with $Z_{mis,1}^{(m)}$ by drawing randomly from the predictive distribution $f(Z_{j,mis}|\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}, \widehat{\theta}^{(m)})$.

These three steps are conducted iteratively until convergence. We obtain the last M imputed data sets for the following analyses. In the third step, instead of fixing one $\widehat{\theta_j}$ for all iterations, we randomly draw $\widehat{\theta_j}^{(m)}$ from the distribution and use it to predict $Z_{j,mis}$ at each iteration. This strategy can guarantee that our imputations are proper (Nielsen, 2003). Details of MICE-IURR for three types of data can be found as Supplementary Method S2 online.

2.3 Simulation Studies

Extensive simulations are conducted to evaluate the performance of the two proposed methods MICE-DURR and MICE-IURR in comparison with the standard MICE and several other existing methods under general missing data patterns. For MICE-DURR and MICE-IURR, we consider three regularization methods, namely, lasso, EN and Alasso. We summarize the simulation results over 200 Monte Carlo (MC) data sets. Following Shah et al. (2014), when applying MI methods, we generate 10 imputed data sets for subsequent analysis, which is our primary goal. To benchmark the bias and loss of efficiency in parameter estimation, two additional approaches that do not involve imputations are also included: a gold standard (GS) method that uses the underlying complete data before missing data are generated, and a complete-case analysis (CC) method that uses only complete-cases for which all the variables are observed(Little and Rubin, 2002).

The setup of the simulations is similar to what was used in Zhao and Long (2013). Specifically, the sample size is fixed at $n = 100$ and each simulated data set includes Y, the fully observed outcome variable, and $Z = (Z_1, \dots, Z_p)$, the set of predictors and auxiliary variables. We consider settings with $p = 200$ and $p = 1000$. We consider Z_1, Z_2, and Z_3 having missing values, which follow a general missing data pattern. We first generate $(Z_4, \dots, Z_p)$ from a multivariate normal distribution with mean $(0, \dots, 0)_{p-4}$ and a first order autoregressive covariance matrix with autocorrelation ρ varying as 0, 0.1, 0.5, and 0.9. Given $(Z_4, \dots, Z_p)$, variables Z_1, Z_2, and Z_3 are generated independently from a normal distribution $N(1 + Z_{\mathcal{S}}\boldsymbol{\alpha}, 4)$, where $\mathcal{S}$ represents the true active set with a cardinality of q. We further consider settings where $q = 4$ and 20, and $\boldsymbol{\alpha} = (1, \ldots, 1)'_4$ for $q = 4$; $\boldsymbol{\alpha} = (0.2, \ldots, 0.2)'_{20}$ for $q = 20$. For $q = 4$ and 20, the corresponding true active set $Z_{\mathcal{S}} = \{Z_4, Z_5, Z_{50}, Z_{51}\}$ and $\{Z_4, \ldots, Z_{13}, Z_{50}, \ldots, Z_{59}\}$. Given Z, the outcome variable Y is generated from $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \boldsymbol{\epsilon}$, where $\beta_i = 1$, and $\boldsymbol{\epsilon} \sim N(0, 6)$ is random noise and independent of Z_i. Missing values are created in Z_1, Z_2, and Z_3 using the following logit models for the corresponding missing indicators, δ_1, δ_2, and δ_3, $\text{logit}(Pr(\delta_1 = 1)) = -1 - Z_4 + 2Z_5 - Y$, $\text{logit}(Pr(\delta_2 = 1)) = -1 - Z_4 + 2Z_{51} - Y$, and $\text{logit}(Pr(\delta_3 = 1)) = -1 - Z_{50} + 2Z_{51} - Y$, resulting in approximately 40% of observations having missing values.

We compare our proposed MICE-DURR and MICE-IURR with a random forest imputation method (MICE-RF)(Shah et al., 2014) and two KNN methods (Liao et al., 2014). For the KNN methods, we use imputations by the nearest variables (KNN-V) and imputations by the nearest subjects (KNN-S) proposed by Liao et al. (2014). When applying MICE-RF, KNN-V, and KNN-S, the R packages returned errors when the incomplete dataset contains large number of variables (i.e. $p = 1000$). As a result, these three methods are only applied to the setting of $p = 200$. In all simulations, for multiple imputations, 10 complete datasets are generated using each method of interest; then the linear regression model is fitted for Y across each imputed data sets for $(Y, Z_1, Z_2, Z_3, Z_4, Z_5)$ and Rubin's rule is applied to obtain $\hat{\boldsymbol{\beta}}$. While for standard MI procedure that cannot be directly used in the cases of $p > n$, we consider one alternative that was used in Zhao and Long (2013): the true active set $\mathcal{S}$ plus Y are used to impute Z_1, Z_2, and Z_3, denoted by MI-true. In practice, MI-true is not accessible since we don't know the true active set. We use the R package `mice` to implement MI-true.

We calculate the following measures to summarize the simulation results for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$: mean bias, mean standard error (SE), Monte Carlo standard deviation (SD), mean square error (MSE) and coverage rate of the 95% confidence interval (CR).

Table A.1, Table A.2, and Table A.3 summarize the results when $\rho = 0.1$, $\rho = 0.5$

and $\rho = 0.9$, respectively. Within each table different methods are compared and the effects of the cardinality of the true active set q and dimension p are evaluated with the correlation ρ fixed. In all scenarios, the complete-case analysis leads to considerably large bias. GS and MI-true that are not accessible in real world provide negligible bias and their CR are close to the nominal level. However, existing machine learning based imputation methods (i.e. MICE-RF, KNN-V and KNN-S) yield considerable biases. MICE-RF ,with a large bias, tends to obtain a large coverage rate close to 1 and thus imprecisely estimate the parameters. KNN-V and KNN-S, on the other hand, predict the missing values only once and have a small CR, likely a result of improper imputation method. MICE-DURR exhibits substantial bias and MSE, thus performing poorly in our settings. There are some differences between our results and those in Zhao and Long (2013), possibly due to the presence of general missing data patterns. For example, MICE-DURR has a poor performance with considerable bias and MSE in our settings. Note that the direct use of regularized regression method in their paper does improve the accuracy of the estimate in their simulation settings where only one variable has missing values. However, our MICE-IURR method indeed achieves a better performance compared with other imputation methods. In all settings, MICE-IURR method using lasso or EN provides relatively small bias and gives similar results compared with MI-true. When $\rho = 0.1$, the biases and MSEs for MICE-RF, KNN-V, and MICE-DURR decrease as q increases, while the performance of KNN-S deteriorates. Nevertheless, MICE-IURR with three regularized methods gives stable results when q changes. When we fix ρ and q, the results of MICE-DURR and MICE-IURR with $p = 200$ are very similar compared with the results with $p = 1000$.

Compared with Table A.1, Table A.2 and Table A.3 show similar patterns on comparisons among the imputation methods. While in Table A.1, among three MICE-IURR algorithms, Alasso underperforms lasso and EN, this situation does not appear in Table A.2 and Table A.3. When $\rho = 0.5$, the biases and MSEs for MICE-IURR using lasso and EN decrease as q increases from 200 to 1000, while these values rise for MICE-IURR using Alasso.

2.4 Data Examples

We illustrate the proposed methods using two data examples.

2.4.1 Georgia stroke registry data

Stroke is the fifth leading cause of death in the United States and a major cause of severe long-term disability. The Georgia Coverdell Acute Stroke Registry (GCASR) program is funded by Centers for Disease Control Paul S. Coverdell National Acute Stroke Registry cooperative agreement to improve the care of acute stroke patients in the pre-hospital and hospital settings. In late 2005, 26 hospitals initially participated in GCASR program and this number increased to 66 in 2013, which covered nearly 80% of acute stroke admissions in Georgia. Intravenous (IV) tissue-plasminogen activator (tPA) improves the outcomes of acute ischemic stroke patients, and brain imaging is a critical step in determining the use of IV tPA. Time plays a significant role in determining patients' eligibility for IV tPA and their prognosis. The American Heart and American Stroke Association and CDC set a goal that hospitals should complete imaging within 25 minutes of patients arrival to a hospital. The objective of this study, thus, is to identify the factors that might be associated with hospital arrival-to-imaging time. GCASR collected data on 86,322 clinically diagnosed acute stroke admissions between 2005 and 2013. The registry has 203 data elements of which 121 (60%) have missing values, attributed to lack of answers, service not provided, poor documentation and data abstraction or ineligibility of a patient to a specific care. The extent of missingness varies from 0.01% to 28.72%.

In this analysis, we consider arrival-to-CT time the outcome and the other 13 variables the predictors. These 13 variables of interest can be classified into two categories: patient-related variables such as age, gender, health insurance, and medical history; pre-hospital-related variables such as EMS notification. Only gender, age and race are fully observed among 13 variables. A CC analysis is conducted which uses only 15% of the original subjects after the removal of incomplete cases. In addition, MI methods are also used. We first remove variables that have missing rate greater than 40% and the remaining variables are used to impute the missing values of partially observed variables that are of interest. After imputations, each imputed datasets of 86,322 subjects are used to fit the regression models separately and results are combined by Robin's rules. We use a straightforward and popular strategy to handle skip pattern: first treat skipped item as missing data and impute them along with other real missing values, then restore the imputed values for skipped items back to skips in the imputed data sets to preserve skip patterns. We apply five MI methods, namely, the MICE method proposed by Buuren and Groothuis-Oudshoorn (2011) (mice), the MI method proposed by Su, Gelman, Hill, Yajima et al. (2011)(mi), the random forest MICE method proposed by Shah et al. (2014) (MICE-RF), and our MICE-DURR and MICE-IURR methods. When applying KNN-V and KNN-S, the R software returned errors. Thus, KNN-V and KNN-S are not included in this data example.

Table 2.1 provides the results from our data analyses. In the CC analysis, only NIH stroke score and race are shown to be associated with the arrival-to-CT time. The results from all five MI methods are similar in terms of the p-value and the direction of the association. By comparison, while only 2 variables are shown to be statistically significant in the CC analysis, this number increases to 11, 11, 10, 9 and 9 for mi, mice, MICE-RF, MICE-DURR, and MICE-IURR, respectively. For example, after adjusting for other variables, the mean arrival-to-CT time in patients that arrive during the day time (Day) was 18.4 minutes shorter than that in patients arriving at night ($p = 0.036$) based on MICE-IURR imputation. Health insurance and three variables about history of diseases become statistically significant after we apply the MI methods. However, NIH stroke score and race, which are shown to be statistically significant by CC analysis, turn out to be not significant by MICE-DURR and MICE-IURR.

2.4.2 Prostate cancer data

The second data set is from a prostate cancer study (GEO GDS3289). It contains 99 samples, including 34 benign epithelium samples and 65 non-benign samples, with 20,000 genomic biomarkers. Missing values are present for 17,893 biomarkers, nearly 89% of all genomic biomarkers in this data set. In this analysis, we consider a binary outcome y, defined as $y = 1$ if it is a benign sample and $y = 0$ if otherwise, and test whether some genomic biomarkers are associated with the outcome. For the purpose of illustration, we choose three biomarkers (FAM178A, IMAGE:813259 and UGP2), for which the missing rates are 31.3%, 45.5% and 26.3%, respectively. We conduct a logistic regression of y on the three biomarkers. In this analysis, *mi* and *mice* packages give error messages and MICE-RF approach is computationally very expensive. Therefore, we only use our two proposed MI methods (MICE-DURR and MICE-IURR) and the KNN-V and KNN-S methods in addition to the complete-case analysis. All 2107 biomarkers that do not have missing values are used to impute missing values in the three biomarkers.

Table 2.2 presents the results on logistic regression for the prostate cancer data. Based on our results, all three biomarkers become statistically significant after using our multiple imputation methods, except in one case that the p-value of UGP2 after MICE-DURR method is slightly larger than 0.05. In addition, in most cases, the estimates and p-values by MICE-DURR are consistent with those results by MICE-IURR. For example, the regression coefficients of biomarker (IMAGE:813259) after using two different multiple imputations (MICE-DURR and MICE-IURR) are 3.47 and 3.50, with p-values of 0.031 and 0.039, respectively.

2.5 Discussion

We investigate two approaches for multiple imputation for general missing data patterns in the presence of high-dimensional data. Our numerical results demonstrate that the MICE-IURR approach performs better than the other imputation methods considered. The MICE-DURR approach, on the other hand, exhibits large bias and MSE. Two data examples further showcase limitations of the existing imputation methods considered.

As alluded to earlier, while MICE is a flexible approach for handling different data types, its theoretical properties are not well-established. The specification of a set of conditional regression models may not be compatible with a joint distribution of the variables being imputed. Liu et al. (2013) established technical conditions for the convergence of the sequential conditional regression approach if the stationary joint distribution exists, which, however, may not happen in practice. Zhu and Raghunathan (2014) assessed theoretical properties of MI for both compatible and incompatible sequences of conditional regression models. However, their results are established for the missing data pattern where each subject may have missing values in at most one variable.

Table 2.1: Regression coefficients estimates of the Georgia stroke registry data. KNN-V and KNN-S are not included because of errors. NPO, nil per os, Latin for "nothing by mouth", a medical instruction to withhold oral intake of food and fluids from a patient. P-value, (); 95% confidence interval, [ ]. The p-values in bold signify a variable that is significant at $\alpha = 0.05$.

Characteristics	CC	mi	mice	MICE-RF	MICE-DURR	MICE-IURR
NIH stroke score	-1.95 (**<0.001**)	-2.04 (**0.010**)	-6.07 (**<0.001**)	-5.01 (**<0.001**)	-1.10 (0.236)	-1.00 (0.176)
	[-2.7,-1.2]	[-3.48,-0.6]	[-8.5,-3.64]	[-6.38,-3.63]	[-3.04,0.84]	[-2.5,0.49]
EMS pre-notification	-3.17 (0.590)	-19.83 (**0.043**)	-0.82 (0.957)	-5.23 (0.604)	-2.22 (0.819)	-5.92 (0.658)
	[-14.69,8.35]	[-38.7,-0.96]	[-34.04,32.4]	[-25.35,14.9]	[-21.59,17.14]	[-34.44,22.59]
Serum total lipid	-0.07 (0.201)	-0.47 (**<0.001**)	-0.52 (**<0.001**)	-0.36 (**<0.001**)	-0.26 (**0.036**)	-0.26 (**0.005**)
	[-0.18,0.04]	[-0.62,-0.32]	[-0.7,-0.33]	[-0.53,-0.19]	[-0.49,-0.02]	[-0.43,-0.08]
Age	0.02 (0.936)	-0.87 (**0.022**)	-0.78 (**0.042**)	-0.71 (0.061)	-0.76 (**0.045**)	-0.80 (**0.037**)
	[-0.51,0.56]	[-1.62,-0.12]	[-1.53,-0.03]	[-1.46,0.03]	[-1.51,-0.02]	[-1.54,-0.05]
Male(referent: female)	5.33 (0.372)	21.20 (**0.013**)	24.83 (**0.004**)	19.15 (**0.025**)	16.57 (0.053)	16.54 (0.053)
	[-6.37,17.02]	[4.41,37.98]	[8,41.66]	[2.41,35.89]	[-0.24,33.38]	[-0.19,33.27]
White(referent: African American)	-16.64 (**0.007**)	-14.44 (0.107)	-20.07 (**0.028**)	-17.64 (**0.048**)	-14.11 (0.114)	-13.85 (0.121)
	[-28.82,-4.45]	[-32.01,3.14]	[-37.98,-2.15]	[-35.16,-0.12]	[-31.62,3.41]	[-31.34,3.65]
Health insurance by medicare	-4.07 (0.617)	-24.95 (**0.032**)	-24.89 (**0.032**)	-24.35 (**0.036**)	-24.36 (**0.036**)	-24.05 (**0.038**)
	[-20.04,11.9]	[-47.72,-2.19]	[-47.7,-2.08]	[-47.13,-1.57]	[-47.11,-1.6]	[-46.8,-1.3]
Arrive in the daytime	4.94 (0.420)	-23.10 (**0.011**)	-24.64 (**0.006**)	-10.27 (0.275)	-18.97 (**0.043**)	-18.41 (**0.045**)
	[-7.07,16.96]	[-40.89,-5.31]	[-42.35,-6.94]	[-28.74,8.2]	[-37.38,-0.56]	[-36.39,-0.42]
NPO	8.37 (0.393)	58.79 (**0.001**)	121.04 (**<0.001**)	81.03 (**0.001**)	39.98 (**0.006**)	43.31 (**0.001**)
	[-10.84,27.58]	[26.65,90.93]	[76.21,165.88]	[39.6,122.45]	[11.87,68.08]	[17.25,69.37]
History of stroke	-2.57 (0.695)	-36.55 (**0.001**)	-34.10 (**0.002**)	-28.99 (**0.009**)	-31.96 (**0.008**)	-33.24 (**0.002**)
	[-15.43,10.29]	[-57.15,-15.96]	[-55.86,-12.35]	[-50.72,-7.27]	[-55.52,-8.41]	[-54.47,-12]
History of TIA	-16.30 (0.097)	-64.53 (**<0.001**)	-89.47 (**<0.001**)	-64.94 (**<0.001**)	-62.39 (**<0.001**)	-60.34 (**<0.001**)
	[-35.54,2.94]	[-95.93,-33.13]	[-123.95,-55]	[-97.47,-32.41]	[-96.59,-28.19]	[-92.47,-28.2]
History of cardiac valve prosthesis	-27.25 (0.349)	89.28 (**0.016**)	136.94 (**<0.001**)	126.22 (**0.022**)	103.18 (**0.008**)	104.15 (**0.007**)
	[-84.27,29.78]	[16.98,161.59]	[79.4,194.48]	[19.67,232.77]	[27.19,179.16]	[28.31,179.98]
Family history of stroke	-17.33 (0.406)	-85.07 (**0.014**)	-51.10 (0.078)	-82.91 (**0.022**)	-79.79 (**0.028**)	-76.82 (**0.034**)
	[-58.18,23.51]	[-153.23,-16.92]	[-107.91,5.72]	[-153.95,-11.87]	[-150.94,-8.65]	[-147.91,-5.74]

Table 2.2: Regression coefficients estimates of the prostate cancer data. MICE-RF is not included because of errors. P-value, (); 95% confidence interval, [ ]. The p-values in bold signify a variable that is significant at $\alpha = 0.05$.

Biomarkers	CC	KNN-V	KNN-S	MICE-DURR	MICE-IURR	Missing-rate
FAM178A	5.80 (0.119)	5.62 (**<0.001**)	5.33 (**<0.001**)	4.43(**0.003**)	4.70(**0.002**)	31.3%
	[-1.49,13.09]	[2.62,8.62]	[2.31,8.35]	[1.61,7.25]	[1.76,7.64]	
IMAGE:813259	6.03 (0.151)	4.20 (**0.009**)	4.43 (**0.016**)	3.47 (**0.031**)	3.50 (**0.039**)	45.5%
	[-2.2,14.26]	[1.06,7.34]	[0.82,8.04]	[0.37,6.57]	[0.23,6.77]	
UGP2	-2.57 (0.386)	-3.44 (**0.021**)	-3.45 (**0.021**)	-2.32 (0.067)	-3.15 (**0.025**)	26.3%
	[-8.37,3.23]	[-6.36,-0.52]	[-6.37,-0.53]	[-4.77,0.13]	[-5.85,-0.45]	

Chapter 3

Privacy-Preserving Methods for Horizontally Partitioned Incomplete Data

3.1 Introduction

The last decade has seen tremendous advances in the amount of data we routinely collect from multiple sources in almost every field. For instance, hospitals have been aggregating electronic health records (EHRs) into medical databases. In addition, the US federal government and other nonprofit organizations have been vastly acquiring health-care knowledge, including data from clinical units and information on patients from insurance companies. In parallel, a number of statistical and data mining methods have been developed to analyze health-care data from multiple sources, aiming to predict epidemics, prevent disease and improve quality of life.

Distributed health data networks are systems that allow secure remote analysis of separate data sets from multiple medical sources (Maro et al., 2009). The networks allow the data to be physically controlled by the data owners, who have the best understanding of the applications to their own data. Such networks also eliminate the need to create and maintain central data repositories. While data cleaning and data analyses are straightforward when the data are collected and stored in a centralized location, such centralization of the data may not be practical for a variety of reasons, including institutional policies and privacy concerns. (Li et al., 2015). For example, Veteran's Health Administration policies require EHR data to remain only at VA's facilities. In addition, improper disclosure of individual-level data has serious implications, such as discriminations for employment, insurance, or education (Naveed et al., 2014). A large body of research has shown that given some background information of an individual, an adversary can learn (from "de-identified" data) sensitive information about the victim (Jiang, Sarwate and Ohno-Machado, 2013; Homer et al., 2008; Brakerski, 2012; Gymrek et al., 2013; Wang et al., 2009). In this chapter, we investigate the situation in which the data are horizontally partitioned. That is, different institutions (sites) have the same characteristics for different individuals. For instance, several local hospitals may want to combine their patients' data to improve the precision of analyses of the general patient population. Due to the aforementioned institutional policy and privacy concerns, institutions are not allowed or willing to fully share their private observations with others, in spite of the fact that they still want to benefit from the collaboration. Under such circumstances, sharing only aggregated statistics among institutions are privacy-preserving and thus acceptable. These existing privacy-preserving algorithms are mainly for the purpose of statistical analysis, such as linear regression and logistic regression, assuming data are complete.

However, due to collection errors and systemic reasons, missing data are frequently encountered in biomedical studies, particularly those requiring data from multiple institutions. Missing data problem reduces the usable sample size and is shown to have significant adverse impact on conclusions drawn from studies such as GWAS (Denny et al., 2013) and computational phenotyping (Newton et al., 2013). Therefore, researchers have devoted a lot of attentions to tackle the challenge of missing data. Before deciding on the best way forward in handling missing data, the pattern and mechanism of missingness should be considered (Penny and Atkinson, 2012). Missing data commonly follow a univariate pattern where the missing values occur on a single variable only, or a general pattern where more than one variables are partially observed. Three missing data mechanisms are introduced by Little and Rubin (2014): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Two popular approaches to addressing the problem of missing data are inverse-probability weighting (IPW) (Höfler et al., 2005) and multiple imputation (MI) (Little and Rubin, 2014). IPW is first formally introduced by Horvitz and Thompson (1952). The idea of IPW is to correct the bias due to the unrepresentativeness of the subgroup of complete cases. This process is realized by weighting each subject of the subgroup by the inverse of the probability of observing a complete case. On the other hand, MI methods replace each missing value multiple times by predicted values drawn from an imputation model. The predictive imputation model is estimated from the observed data, which contain no missing values. In this way, the extra uncertainty is reflected due to the fact that the regression parameters can be estimated, but not determined, from the observed data. After obtaining multiple complete datasets, each dataset is analyzed and an estimate of the analysis model parameters, $\boldsymbol{\theta}$, is calculated. Rubin (1987) proposes rules to combine these estimates and further calculate the variance of the combined estimate. In the presence of general missing data patterns, multiple imputation by chained equations (MICE) method is widely adopted and has been shown to achieve superior performance in practice (Raghunathan and Siscovick, 1996; Buuren and Groothuis-Oudshoorn, 2011).

Under MAR, a naïve MI approach for horizontally partitioned incomplete data is to conduct MI within each institution and then perform the distributed analysis. This approach has two limitations. First, it will lead to large variability in imputation and subsequent analysis. Second, when one variable is missing for all observations in a single institution, this variable cannot be imputed in that institution using this naïve MI approach. Jagannathan and Wright (2008) proposed a privacy-preserving lazy decision-tree imputation algorithm for data that are horizontally partitioned between two sources. Unfortunately, their complex decision trees may overfit the data and become unstable. Moreover, their algorithm cannot be applied to general missing data patterns and the multiple sources case.

In this chapter, we develop a framework for handling missing data under the distributed environment. That is, each institution keeps their own local data in local private zone and calculates aggregated statistics (e.g., co-variance, kernel matrix, etc.) that are necessary for handling missing data. These statistics are stored in the shared zone and will be exchanged across all institutions to build a global missing data model. In other words, during the collaboration among institutions, information exchange occurs only in the shared zone of each institution and no individual-level data will be transmitted to protect the privacy. Figure 3.1 presents the conceptual architecture of the framework for privacy-preserving methods for missing data. The missing data models we propose in this chapter include: 1) privacy-preserving IPW and MI for horizontally partitioned data assuming MAR, with univariate missing data patterns and 2) a privacy-preserving MICE for horizontally partitioned data assuming MAR, with general missing data patterns. We further investigate a re-weighting privacy-preserving MI for horizontally partitioned data that are MNAR. To the best of our knowledge, this is the first work of investigating privacy-preserving methods for horizontally partitioned incomplete data from multiple sources, in distributed environments.

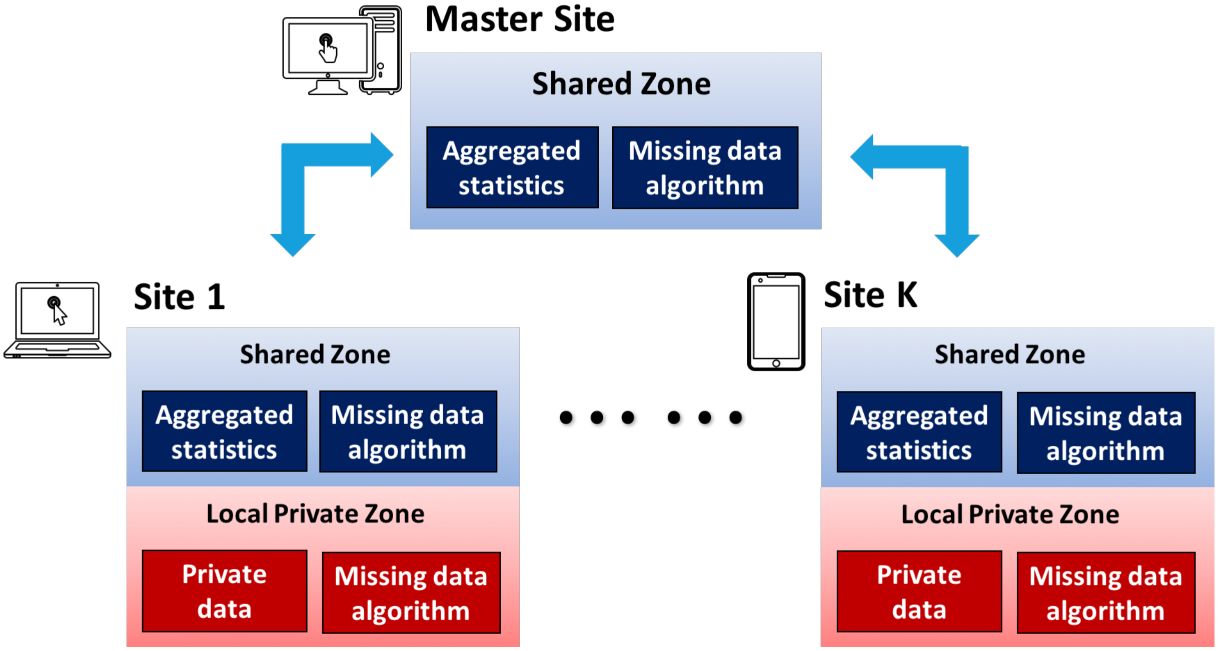


Figure 3.1: Conceptual architecture of the privacy-preserving framework for handling missing data

3.2 Methodology

We consider a multiple linear model for the regression of outcome Y on p covariates $X_1, ..., X_p$. Let $X = (\mathbb{1}, X_1, ..., X_p)$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)$ denote the model parameters of interest such that

$$Y = \theta_0 + \theta_1 X_1 + ... + \theta_p X_p + \epsilon, \tag{3.1}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, $Y = (y_1, ..., y_n)^T$, $X_j = (x_{1,j}, ..., x_{n,j})^T$ and n is the total number of individuals. This model is referred to as the "analysis model" throughout this chapter. The values for individual i $(i = 1, ..., n)$ are $x_{i,1}, x_{i,2}, ..., x_{i,p}$, and y_i. In this section, we discuss horizontally partitioned data with K institutions, which share the same covariates (features) of exclusive individuals. We refer to Y and X, respectively, as the "pooled" outcome vector and the "pooled" design matrix, such that

$$Y = \begin{pmatrix} Y^{site_1} \\ \vdots \\ Y^{site_K} \end{pmatrix}, \quad X = \begin{pmatrix} X^{site_1} \\ \vdots \\ X^{site_K} \end{pmatrix},$$

where Y^{site_k} and X^{site_k} $(1 \leq k \leq K)$ are data from institution k with n_k subjects.

Note that each X^{site_k} is an $n_k \times (p+1)$ matrix with the first column of 1's and we let $n = \sum_{k=1}^{K} n_k$ be the total number of individuals for the "pooled" data.

The privacy-preserving methods for horizontally partitioned data with missingness are described in Section 3.2.1, Section 3.2.2 and Section 3.2.3. For the ease of exposition, we let $p = 2$ and consider a univariate missing data pattern where only one variable, X_1, has missing values while other variables are fully observed. Let $r_i = 1$ if $x_{i,1}$ is observed, and $r_i = 0$ otherwise, for $i = 1, ..., n$. Let $R = (r_1, ..., r_n)^T$ and $R = (R^{site_1 T}, ..., R^{site_K T})^T$, where R^{site_K} is the indicator vector that belongs to institution k $(k = 1, ..., K)$.

3.2.1 Privacy-preserving inverse probability weighting for horizontally partitioned data

If data are complete, $\boldsymbol{\theta}$ is estimated as the value $\hat{\boldsymbol{\theta}}$ that minimizes $\sum_{i=1}^{n} U_i(\boldsymbol{\theta})$, where $U_i(\boldsymbol{\theta}) = (y_i - \theta_0 - \theta_1 x_{i,1} - \theta_p x_{i,2})^2$. In the presence of missing data, one common method is to estimate $\boldsymbol{\theta}$ using only complete cases. This is known as a complete case (CC) analysis. The CC estimator is consistent under MCAR mechanism. However, it will generally not be consistent in many other situations (Seaman and White, 2013). When the data are MAR, an alternative approach is to fit the model ignoring incomplete cases as well, but more weight is given to some complete cases than others. The weight w_i is the inverse of the probability $p_i = Pr(r_i = 1)$, which is the probability of the individual i being a complete case. This inverse probability weighting (IPW) approach minimizes a different objective function. That is, the IPW estimator is the value that minimizes $\sum_{i=1}^{n} r_i w_i U_i(\boldsymbol{\theta})$.

A key step for IPW is to build a logistic regression model to predict p_i based on Y and the fully observed variable X_2. Let $\boldsymbol{Z} = (\mathbb{1}, Y, X_2)$. The logistic regression becomes $\text{logit}(Pr(R = 1)) = \boldsymbol{Z\beta}$. The maximum likelihood estimator (MLE) of the logistic regression can be obtained by utilizing Newton Raphson optimization: $\hat{\boldsymbol{\beta}}_{MLE} = \arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i log(p_i) + (1 - r_i) log(1 - p_i)$, where $p_i = 1/(1 + exp(-(1, y_i, x_{i,2}) \times \boldsymbol{\beta})$. We update the estimator iteratively until it convergences to a stationary point $\hat{\boldsymbol{\beta}}_{MLE}$. The updating step is as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{new} &= \hat{\boldsymbol{\beta}}^{old} - l''(\hat{\boldsymbol{\beta}}^{old})^{-1} l'(\hat{\boldsymbol{\beta}}^{old}) \\ &= \hat{\boldsymbol{\beta}}^{old} + (\boldsymbol{Z}^T \boldsymbol{W}^{old} \boldsymbol{Z})^{-1} \boldsymbol{Z}^T (R - P^{old}),\end{aligned} \tag{3.2}$$

where $\boldsymbol{W}^{old} = \begin{pmatrix} p_1^{old}(1 - p_1^{old}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n^{old}(1 - p_n^{old}) \end{pmatrix}$, $P^{old} = (p_1^{old}, ..., p_n^{old})^T$, and $p_i^{old} = 1/(1 + exp(-(1, y_i, x_{i,2}, ..., x_{i,p}) \times \boldsymbol{\beta}^{old})$.

For horizontally partitioned data as we focus in this chapter, since we may not have the "pooled" data (i.e. $\boldsymbol{Z}$, $\boldsymbol{W}$ and R), we propose a privacy-preserving IPW (PPIPW-H) method, in which a distributed Newton Raphson optimization (Wu et al., 2012) is applied and only intermediate statistics from each institutions are exchanged across institutions. Let $\boldsymbol{Z}^{site_k}$ be the design matrix that belongs to institution k ($k = 1, ..., K$). $\boldsymbol{W}^{site_k old}$ is the diagonal matrix partitioning from $\boldsymbol{W}^{old}$, corresponding to the k-th institution. Likewise,

$P^{site_k old}$ is the vector partitioning from P^{old}, corresponding to the k-th institution. Since $\boldsymbol{Z}^T\boldsymbol{W}^{old}\boldsymbol{Z} = (\boldsymbol{Z}^{site_1})^T\boldsymbol{W}^{site_1 old}\boldsymbol{Z}^{site_1} + \cdots + (\boldsymbol{Z}^{site_K})^T\boldsymbol{W}^{site_K old}\boldsymbol{Z}^{site_K}$ and $\boldsymbol{Z}^T(R - P^{old}) = (\boldsymbol{Z}^{site_1})^T(R^{site_1} - P^{site_1 old}) + \cdots + (\boldsymbol{Z}^{site_K})^T(R^{site_K} - P^{site_K old})$. The updating step of the distributed Newton Raphson algorithm becomes:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{new} = \hat{\boldsymbol{\beta}}^{old} + \{(\boldsymbol{Z}^{site_1})^T\boldsymbol{W}^{site_1 old}\boldsymbol{Z}^{site_1} + \cdots + (\boldsymbol{Z}^{site_K})^T\boldsymbol{W}^{site_K old}\boldsymbol{Z}^{site_K}\}^{-1}\cdot \\ \{(\boldsymbol{Z}^{site_1})^T(R^{site_1} - P^{site_1 old}) + \cdots + (\boldsymbol{Z}^{site_K})^T(R^{site_K} - P^{site_K old})\},\end{aligned} \tag{3.3}$$

Instead of merging the data from multiple sources to calculate $\boldsymbol{Z}^T\boldsymbol{W}^{old}\boldsymbol{Z}$ and $\boldsymbol{Z}^T(R - P^{old})$, PPIPW-H lets each institution calculate the aggregated statistics (i.e. $(\boldsymbol{Z}^{site_k})^T\boldsymbol{W}^{site_k old}\boldsymbol{Z}^{site_k}$ and $(\boldsymbol{Z}^{site_k})^T(R^{site_k} - P^{site_k old})$) first based on the local private data, following by the summation step of these statistics. PPIPW-H only leverages locally aggregated statistics and thus is privacy-preserving. More importantly, since Equation 3.2 and Equation 3.3 are essentially calculating the same $\hat{\boldsymbol{\beta}}^{new}$, PPIPW-H using the distributed Newton Raphson algorithm will provide the same $\hat{\beta}_{MLE}$ as if it were constructed using the "pooled" data.

The estimated weight $\hat{w}_i = 1/\hat{p}_i$, where $\hat{p}_i$ is the predicted probability built upon $\hat{\boldsymbol{\beta}}_{MLE}$. Let $\boldsymbol{w}_k$ be the vector of weights of individuals from institution k ($k = 1, ..., K$). Then, PPIPW-H solves $\hat{\boldsymbol{\theta}}_{IPW} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} r_i\hat{w}_iU_i(\boldsymbol{\theta})$, from a weighted linear regression, using only intermediate statistics shared by each institution and obtains

$$\hat{\boldsymbol{\theta}}_{IPW} = (\sum_{k=1}^{K}(X^{site_k})^T\boldsymbol{V}^{site_k}X^{site_k})^{-1}(\sum_{k=1}^{K}(X^{site_k})^T\boldsymbol{V}^{site_k}Y^{site_k}), \tag{3.4}$$

where $\boldsymbol{V}^{site_k} = diag(R^{site_k} \circ \boldsymbol{w}^{site_k})$ and "$\circ$" stands for the element-wise products of two vectors. Using this distributed linear regression, Equation 3.4 demonstrates itself to be an additional step of protecting privacy without sharing individual-level data. We then resample the original data set repeatedly and apply the above procedure to get the estimate $\hat{\boldsymbol{\theta}}_{IPW}^{(b)}$ for the bootstrap sample. We use the sample standard variance of $\hat{\boldsymbol{\theta}}_{IPW}^{(b)}$ to estimate the standard error of $\hat{\boldsymbol{\theta}}_{IPW}$.

3.2.2 Privacy-preserving multiple imputation for horizontally partitioned data of univariate missing data patterns

Rather than only using complete cases, multiple imputation method replaces each missing data point with a set of possibles values (M times) from its predictive distribution given observed data. MI has been arguably the most popular method for handling missing data in practice. MI accounts for the uncertainty of imputation by generating multiply imputed datasets. Subsequently, each imputed dataset is analyzed separately using the standard "analysis model". The results across all imputed datasets are then combined following Rubin's rule.

To illustrate the standard and our proposed privacy-preserving method for MI, let's still consider the univariate missing data pattern with 2 covariates in the "analysis model". Suppose that the partially observed variable X_1 is continuous. The standard MI method first fits the imputation model $X_1 = \alpha_0 + \alpha_1 Y + \alpha_2 X_2 + \zeta$, where $\zeta \sim \mathrm{N}(0, \tau^2)$, using the complete cases and obtains the parameter estimate $\hat{\boldsymbol{\alpha}}_{MLE}$ and its variance $\hat{V}_{\boldsymbol{\alpha}}$. Let $X_{1,obs}$ be the observed component of X_1 and $\boldsymbol{Z}$ be the design matrix in the imputation model. Then $\hat{\boldsymbol{\alpha}}_{MLE} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T X_{1,obs}$. To protect the privacy, we propose a privacy-preserving MI (PPMI-H) that can still obtain the same parameter estimate and the variance, without using the pooled data (i.e. $\boldsymbol{Z}$ and $X_{1,obs}$). PPMI-H is also built under the proposed framework that only intermediate statistics are calculate and transmitted in the shared zone. Let $\boldsymbol{Z}^{site_k}$ and $X_{1,obs}^{site_k}$ be the k-th partitions of $\boldsymbol{Z}$ and $X_{1,obs}$, corresponding to institution k. PPMI-H fits a distributed linear regression to estimate the parameters of the imputation model and stores them in the shared zone to be accessible to all institutions.

The MLE of PPMI-H is as follows,

$$\hat{\boldsymbol{\alpha}}_{MLE} = \{(\boldsymbol{Z}^{site_1})^T\boldsymbol{Z}^{site_1}+\cdots+(\boldsymbol{Z}^{site_K})^T\boldsymbol{Z}^{site_K}\}^{-1}\{(\boldsymbol{Z}^{site_1})^T X_{1,obs}^{site_1}+\cdots+(\boldsymbol{Z}^{site_K})^T X_{1,obs}^{site_K}\}. \tag{3.5}$$

Of note, in Equation 3.5, each institution calculates the aggregated statistics (i.e. $(\boldsymbol{Z}^{site_k})^T\boldsymbol{Z}^{site_k}$ and $(\boldsymbol{Z}^{site_k})^T X_{1,obs}^{site_k}$) using local data in private zones and then share them across institutions in order to obtain $\hat{\boldsymbol{\alpha}}_{MLE}$. Then we send $\hat{\boldsymbol{\alpha}}_{MLE}$ back to local sites for the parallel computing of the residual sum of squares (RSS). That is, RSS $= \sum_{k=1}^{K} \text{RSS}^{site_k} = \sum_{k=1}^{K} ||X_{1,obs}^{site_k} - \boldsymbol{Z}^{site_k}\hat{\boldsymbol{\alpha}}_{MLE}||^2$. Consequently, the master site can calculate the estimate of the variance of $\hat{\boldsymbol{\alpha}}_{MLE}$, as $\widehat{\text{Var}(\hat{\boldsymbol{\alpha}}_{MLE})} = \frac{\text{RSS}}{\sum r_i - 3}\{\sum(\boldsymbol{Z}^{site_k})^T\boldsymbol{Z}^{site_k}\}^{-1}$. After the imputation model is built and parameter estimate as well as its variance are obtained, PPMI-H draws $\hat{\boldsymbol{\alpha}}$, M times from its distribution $\text{N}(\hat{\boldsymbol{\alpha}}_{MLE}, \widehat{\text{Var}(\hat{\boldsymbol{\alpha}}_{MLE})})$ and obtain $\hat{\boldsymbol{\alpha}}^{(1)}, \cdots, \hat{\boldsymbol{\alpha}}^{(M)}$. We also draws $\hat{\tau}$ from its posterior distribution M times and get $\hat{\tau}^{(1)}, \cdots, \hat{\tau}^{(M)}$. For PPMI-H using horizontally partitioned data, we store these M estimates in the shared zone such that all institutions can take advantage of them in the subsequent step as follows: each institution replaces their missing values M times with predicted values drawn from $\text{N}(\boldsymbol{Z}_{mis}^{site_k}\hat{\boldsymbol{\alpha}}^{(m)}, \hat{\tau}^{(m)2})$, where $\boldsymbol{Z}_{mis}^{site_k} = (\mathbb{1}, Y_{mis}^{site_k}, X_{1,mis}^{site_k})$. Equation 3.5 demonstrates that PPMI-H leverages only aggregated statistics but resulting the same estimates of the imputation model as if them were obtained using the "pooled" data. With M imputed data sets generated by the aforementioned steps, the standard MI would fit the "analysis model" (i.e. $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \epsilon$)using each imputed dataset and then combined the results by Rubin's rules to get the final estimate. Similarly, our PPMI-H fits the same "analysis model" using a single dataset but under our proposed framework. Let $X^{site_k(m)}$ be the design matrix of institution k with the missing values replaced by the m-th imputation. PPMI-H estimates the parameters of the "analysis model" using a single imputed dataset by $\hat{\boldsymbol{\theta}}^{(m)} = (X^{site_1(m)T}X^{site_1(m)}+\cdots X^{site_K(m)T}X^{site_K(m)})^{-1}(X^{site_1(m)T}Y^{site_1}+\cdots X^{site_K(m)T}Y^{site_K})$. Specifically, $X^{site_k(m)T}X^{site_k(m)}$ and $X^{site_k(m)T}Y^{site_k}$ are calculated in private zones while $\hat{\boldsymbol{\theta}}^{(m)}$ is shared across institutions. Bootstrap method is applied to obtain the estimate of the standard error of $\hat{\boldsymbol{\theta}}^{(m)}$. After obtaining the results from the master site, we apply Rubin's rules to combine them and achieve the final result.

Adopting a similar strategy, we can extend the PPMI-H method for partially observed variable with different types. For example, if X_1 is a binary variable, Equation (3.5) can be replaced with a distributed logistic regression with the similar idea of using the distributed Newton Raphson optimization, as we show in Section 3.2.1.

3.2.3 Privacy-preserving multiple imputation for general missing data patterns

Section 3.2.1 and Section 3.2.2 focus on the univariate missing data patterns that only one variable has missing values, i.e. X_1 is partially observed. However, values of multiple variables can be missing for some reason and this situation is often encountered in biomedical research. We call these general missing data patterns and we discuss the distributed method of handling them.

Without loss of generality, we assume that the first l ($l \leq q$) covariates, i.e. $(X_1, ..., X_l)$, contain missing values. Let $X_{j,obs}$ and $X_{j,mis}$ denote the observed and missing components for X_j ($j = 1, ..., l$). We denote the collection of the outcome and the $p - 1$ covariates except X_j by X_{-j}, i.e. $X_{-j} = (Y, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)$. We assume the hypothetically complete data $(Y, X_1, ..., X_p)$ are partially observed random draw from a multivariate distribution $f(Y, X_1, ..., X_p|\boldsymbol{\alpha})$ that is specified by the unknown parameters $\boldsymbol{\alpha}$. The standard MICE algorithm obtains the posterior distribution of $\boldsymbol{\alpha}$ by sampling iteratively from the conditional distributions of partially observed variables with the form: $f(X_1|X_{-1}, \boldsymbol{\alpha_1}), ..., f(X_l|X_{-l}, \boldsymbol{\alpha_l})$.

We propose a privacy-preserving MICE (PPMICE) that starts with a simple imputation for every missing value. For example, each missing value can replaced by the mean of that variable. For a distributed environment, the mean can be calculated as the average value of the "local" means weighted by the sample size of every institute. It is the same as the from the "pooled" data. We denote $(X_1^{(0)}, ..., X_l^{(0)})$ the initial values of $(X_1, ..., X_l)$ with missing values filled in. In order to obtain M imputed datasets at the end, we first replicate the initial values M times and denote them by $(X_1^{(0,m)}, ..., X_l^{(0,m)})$, where $m = 1, ..., M$. At iteration t $(t \geq 1)$, for the m-th imputation, given data $X_1^{(t-1,m)}, ..., X_l^{(t-1,m)}$ from the previous iteration, the new parameter estimate $\widehat{\boldsymbol{\alpha}}_j^{(t,m)}$ for variable j is generated from $f(\boldsymbol{\alpha}_j | X_{j,obs}, Y, X_1^{(t,m)}, ..., X_{j-1}^{(t,m)}, X_{j+1}^{(t-1,m)}, ..., X_l^{(t-1,m)}, X_{l+1}, ..., X_p)$, using PPMI-H. Subsequently, the missing values $X_{j,mis}$ for X_j are replaced with predicted values from the regression model with model parameter $\widehat{\boldsymbol{\alpha}}_j^{(t,m)}$. The iteration entails cycling through imputing X_1 to X_l and the complete dataset at the end of the cycle is used as the initial values for the next iteration.

3.2.4 Sensitivity Analysis under MNAR assumption

All above-mentioned methods are established assuming the missing data are MAR. However, MAR assumption is unlikely to be true in practice, since the probability of non-response might depend on the unseen data themselves. Standard methods for data that are MNAR are complicated as they need to include a model for the reason for dropout. Carpenter et al. (2007) propose a simple approach for approximate analysis under the MNAR assumption using multiple imputations created assuming MAR. The principle of this approach is to re-weight the parameters estimated from the imputed data sets of assuming MAR, so that the weighted parameter estimates represent the true underlying distributions under a MNAR mechanism. They assume that the response variable Y is subject to missing values and the covariances are fully observed, which are different from our setting. In this section, we propose a privacy-preserving re-weighting multiple imputation method and apply it to a more general data setting that has been used throughout this dissertation: data with fully observed response but partially observed covariates.

For simplicity, suppose we have a response variable Y and covariates X_1 and X_2 for

n subjects, where X_1 has missing values. Note that the method can be directly applied to other data settings of univariate missing-data patterns with more than two covariates. The "analysis" model is

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \epsilon,$$

where ϵ follows a normal distribution. Now suppose the first n_1 subjects are missing X_1 and let the observed indicator $R_i = 0$ $(i \leq n_1)$ for these n_1 subjects and $R_i = 1$ for the remaining subjects. Assume that the probability of non-response depends on Y, X_1 and X_2 through the logistic model

$$\text{logit}(Pr(R = 1)) = h(Y, X_2) + \delta X_1 \tag{3.6}$$

, where δ is the sensitivity parameter and needs to be pre-specified and $h(\cdot)$ is an unspecified function of Y and X_2. The data subject to missing values through model (3.6) are considered having a general MNAR mechanism. Let θ be a vector of parameter of interest (i.e., $\theta = (\theta_0, \theta_1, \theta_2)^T$). Assuming MAR, suppose we have imputed $X_{1,mis}$ by $X_{1,mis}^{(1)}, ..., X_{1,mis}^{(M)}$ from $f(X_1|Y, X_2, R = 1)$. Our goal is to put some weights $(w_1, ..., w_M)$ to $X_{1,mis}^{(1)}, ..., X_{1,mis}^{(M)}$ so that they can be viewed as samples from $f(X_1|Y, X_2, R = 0)$. Given the M imputed data sets, θ can be estimated by combining $\hat{\theta}_1 = \hat{\theta}(Y, X_{1,obs}, X_{1,mis}^{(1)}, X_2), ..., \hat{\theta}_M = \hat{\theta}(Y, X_{1,obs}, X_{1,mis}^{(M)}, X_2)$ through those weights:

$$\hat{\theta}_{MNAR} = \sum_{m=1}^{M} w_m \hat{\theta}_m \tag{3.7}$$

Similarly, the estimate of variance is approximated by $V_{MNAR} \approx \tilde{V}_W + (1 + 1/M)\tilde{V}_B$, where

$$\tilde{V}_W = \sum_{m=1}^{M} w_m \hat{\sigma}_m^2, \quad \tilde{V}_B = \sum_{m=1}^{M} w_m (\hat{\theta}_m - \hat{\theta}_{MNAR})^2 \tag{3.8}$$

We use importance sampling to calculate the weights. The weight for samples from

$f(X_1|Y, X_2, R = 0)$ to obtain $f(X_1|Y, X_2, R = 1)$ is simply the ratio

$$\begin{aligned}\frac{f(X_1|Y, X_2, R = 0)}{f(X_1|Y, X_2, R = 1)} &= \frac{f(Y, X_2, R = 0|X_1)f(Y, X_2, R = 1)}{f(Y, X_2, R = 1|X_1)f(Y, X_2, R = 0)} \\ &= \frac{f(R = 1|X_1, X_2, Y)}{f(R = 0|X_1, X_2, Y)} \times \frac{f(Y, X_2, R = 1)}{f(Y, X_2, R = 0)} \\ &= e^{-(h(Y,X_2)+\delta X_1)} \times \frac{f(Y, X_2, R = 1)}{f(Y, X_2, R = 0)} \\ &= e^{-\delta X_1} \times g(Y, X_2, R)\end{aligned} \tag{3.9}$$

Since the term $g(Y, X_2, R)$ is common to all weights and we have n_1 missing values of X_1 $(x_{1,1}, x_{1,2}, ..., x_{1,n_1})$ need to be imputed in the mth imputation, the normalized weight for the mth imputation is proportional to $e^{-\delta \sum_{i=1}^{n_1} x_{1,i}^{(m)}}$, that is

$$w_m \propto e^{-\delta \sum_{i=1}^{n_1} x_{1,i}^{(m)}}, \tag{3.10}$$

where $x_{1,i}^{(m)}$ is the ith element of $X_{1,mis}^{(m)}$. Note that these weights are simple and can be easily calculated from local by adding up the summation of imputed values from each institutions, without sharing each patient' information. That is, $\sum_{i=1}^{n_1} x_{1,i}^{(m)} = \sum_{k=1}^{K} \sum_{\{i \in site_k : i < n_1\}} x_{1,i}^{(m)}$. We denote our privacy-preserving re-weighted MI method for data that are MNAR by PPMI-RW.

3.3 Simulation Studies

3.3.1 Simulation Study when Data are MAR

In this section, we provide three simulation studies to demonstrate our privacy-preserving methods in comparison with the standard methods of addressing missing data under univariate and general missing data patterns. As in Section 3.2, we consider a linear regression model (3.1) as the "analysis model". The simulation results over 1000 Monte Carlo(MC) data sets. Each simulation study has a different way to generate the MC data sets.

In the first study, we explore approaches for addressing continuous variable with missing values under univariate missing data patterns, i.e. only X_1 has missing values. Data $X_1, ..., X_p$ and Y are generated for $n = 200$ and $n = 1000$ individuals. We consider a setting with $p = 2$ in this study. For each individual, X_2 is first generated from a uniform distribution $U(-1, 1)$. Given X_2, variable X_1 is sampled from a normal distribution with variance $\sigma^2_{X_1} = 1$ and mean $\mu_{X_1} = X_2$. Outcome Y is generated from $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and all $\theta_j = 1$ $(j = 0, 1, 2)$. Variable X_1 is missing with probability $\{1 + \exp(1.6 - Y - X_2)\}^{-1}$, resulting in approximately 42% of individuals having missing values. In order to illustrate that our privacy-preserving methods can be applied to binary variable with missing values, we make a little change to the previous data-generating mechanism. Given X_2, instead of sampling X_1 from a normal distribution, we generate X_1 from a Bernoulli distribution with probability $\{1 + \exp(-1 - X_2)\}^{-1}$. The rest procedures are the same.

The third study is to test the performances of methods under general missing data patterns. In this data-generating mechanism, suppose $p = 5$, we assume that X_1, X_2, and X_3 have missing values among $n = 200$ or $n = 1000$ individuals. Fully observed variables X_4 and X_5 are independent and identically distributed $\mathrm{N}(0, 1)$. Given $X_4, ..., X_p$, variables X_1, X_2, and X_3 are generated independently from a normal distribution $\mathrm{N}(1 + X_4 + X_5, 1)$. We use the same way to generate the outcome Y as we describe in the first study. Missing values are created in X_1, X_2, and X_3 using the following logit models for the corresponding missing indicators, δ_1, δ_2, and δ_3, $\mathrm{logit}(Pr(\delta_1 = 1)) = 3 - 0.8Y - 0.1X_4 - 0.2X_5$, $\mathrm{logit}(Pr(\delta_2 = 1)) = 3 - 0.4Y - 0.2X_4 - 0.4X_5$, and $\mathrm{logit}(Pr(\delta_3 = 1)) = 2 - 0.3Y - 0.4X_4 - 0.3X_5$, resulting in approximately 40% of individuals having missing values.

After the data set is generated, we horizontally partition it to mimic a distributed environment where data are stored across several institutes and they can not be pooled because of some privacy issue. We test the performance of our methods with the number of institutes $K = 5$ and $K = 20$, where each institute has the same number of individuals, in other words, n individuals are equally partitioned into K institutes.

For the first two simulation studies that are under the assumption of univariate missing data patterns, we compare our proposed PPIPW-H and PPMI-H with the aforementioned naïve MI (MI-naïve) that conducts MI within each institution and the MI method using pooled data (MI-pooled). For the third study, we do not test PPIPW-H since it is not applicable to the general missing patterns. Note that for MI-naïve, standard MI may not be suitable in a institution that does not have any observed values for some variable. In this case, we replace the missing values for that variable in the institution with the mean based on the observed values of that variable from all the institutions. We generate 100 imputed datasets using each MI method; then a distributed analysis is conducted to fit the "analysis model" in each imputed dataset and Rubin's rule is applied to obtain $\widehat{\boldsymbol{\theta}}$ and their standard errors. We include two approaches that do not involve imputations as benchmarks to evaluate bias and loss of efficiency in parameter estimation: a gold standard (GS) method that utilizes the underlying complete data before missing data are generated, and a complete-case (CC) method that uses only complete-cases for which all variables are observed. We summarize the simulation results of $\hat{\theta}_0$, $\hat{\theta}_1$, and $\hat{\theta}_2$ for the first two studies and the simulation results of $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ for the third study.

Table 3.1 and Table 3.2 show the mean relative bias (RBias), mean standard error (SE), MC standard deviation (SD), mean square error (MSE) and coverage rate of 95% confidence interval (CR) for three parameters and eight analysis methods under univariate missing data patterns. It can be seen that in all scenarios, CC analysis that ignoring incomplete cases yields a strong bias and MSE, indicating the necessary of handling missing data. PPIPW-H which again only includes complete cases in the analysis model, but weights are adopted to rebalance the set of complete cases so that it is representative of the whole sample (Seaman et al., 2012). From the results, PPIPW-H outperforms

CC to some extent in terms of RBias and MSE, but has large SE compared with other methods. This is because PPIPW-H accounts for the uncertainty of the estimated weights. In practice, MI methods are ofter preferred to IPW methods (i.e. PPIPW-H), as it is more efficient. However, MI-naïve method yields a non-ignorable RBias and MSE when the sample size n is small and the number of institutions K is large. The reason is that some institution may have few or no observed values to predict the missing values. For example, when $n = 200$ and $K = 20$, each institution has 10 individuals. We may not observed any values of X_1 from that 10 individuals in a institution and MI-naïve does not have any information to conduct the MI locally in this case. PPMI-H, which analyzes the imputation model in a distributed environment, is proved to work well in this situation. It still preserves the privacy by not sharing the information of each individual as illustrated in Section 3.2.2. It can been seen from the tables that PPMI-H is approximately unbiased by exhibiting small to negligible RBias and performs similarly to MI-pooled, as expected from Section 3.2.2. Even though MI-pooled is more computational efficiency than other imputation methods including PPMI-H, it is not accessible in a distributed environment that we describe to preserve privacy. This is because MI-pooled requires the data to be pooled before we address the missing data.

Table 3.3 summarizes the results for general missing data patterns that X_1, X_2, and X_3 have missing values. The conclusion matches those from Table 3.1 and Table 3.2. First, MI-naïve performs poorly with substantial bias, and this situation deteriorates as K increases. Second, PPMICE is efficient unbiased and has similar results to MI-pooled in regardless of K.

3.3.2 Simulation Study when Data are MNAR

Here we describe a simulation study to evaluate the PPMI-RW method for data that are MNAR. We set $n = 1000$ and X_1, X_2, Y are generated from

$$\begin{pmatrix} X_1 \\ X_2 \\ \epsilon \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.2 & 0.1 & 0 \\ 0.1 & 0.1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right]$$

$$Y = X_1 + X_2 + \epsilon$$

Missing values are created in X_1 using the logit model: $\text{logit}(Pr(R = 0)) = 1 + Y + X_1 + 0.1X_2$, resulting in approximately 37% of subjects having missing values. We perform 200 simulations to generate 200 MC data sets and averaged the resulting estimates of $\theta_0, \theta_1, \theta_2$ (true values: 0, 1, 1). We use $M = 100$ imputations as suggested in Carpenter et al. (2007) and also not burdensome in our simulation setting. We compare PPMI-RW method with GS, CC, MI-naïve, MI-pooled, and MI-pooled-RW. Note that MI-naïve and MI-pooled assume the data are MAR and MI-pooled-RW method re-weight the imputations from MI-pooled assuming the data can be pooled and do not have any privacy constrains.

Table 3.4 summarizes the results. MI-pooled-RW and PPMI-RW remove substantially more of the bias than MI-pooled. Moreover, the MSE decreases dramatically after using the re-weighting technique.

3.4 Data Example

The Georgia Coverdell Acute Stoke Registry (GCASR) data consists of 86,322 patients with clinically diagnosed acute stroke between 2005 and 2013. Data were collected from 61 hospitals in Georgia. A total of 203 characteristics are included in the data, with 60% of which have missing values due to lack of answers, service not provided, poor documentation and data abstraction or ineligibility of a patient to a specific care.

Deng et al. (2016) investigate the effect of 13 characteristics on arrival-to-CT time. These 13 characteristics are patient-related such as age, gender or pre-hospital-related such as EMS notification. With in these features of interest, only gender, age and race are fully observed. In their study, MI methods using the pooled data were performed and 20 complete datasets were produced. In order to protect the privacy of the data, we assume that patient-level data sharing is forbidden and only summary statistics can be possibly exchanged between 61 hospitals. A CC analysis is conducted by removing patients with missing values from each hospital. After the removal, it only consists 15% of the original patients that can subsequently be used to a distributed linear regression. We also consider MI-naïve that each hospital conducts MICE locally. Of note, some hospitals have patients less than 10 which indicates the inefficiency of MI-naïve. This is because no promising values can be predicted from the hospital with most of its patients' information not observed. However, PPMICE can preserve the privacy and make full use of each patient's information as well. We also apply MI-pooled which is not applicable in the distributed environment assumption, to benchmark our PPMICE.

Table 3.5 shows the estimates, p-values, and 95% confidence intervals of each characteristic of interest for four methods. The results for CC and MI-naïve deviate away noticeably from that for MI-pooled. As can be seen, the CC analysis shows only NIH stroke (p<0.001) and race ($p = 0.007$) are associated with the arrival-to-CT time. For MI-naïve, 6 variables have p-values less than .05. However, if we break the privacy-preserving restriction by combining the data from 61 hospitals, the corresponding analysis (MI-pooled) on the joint data shows that 10 variables are statistically significant. Even though 6 statistically significant variables detected by MI-naïve are also found to be associated with the outcome using the joint data (MI-pooled), their estimates are different. This is not the case for our PPMICE. It shows very little difference to the results of MI-pooled in terms of both p-values and estimates. This data example illustrates that PPMICE performs as well as MI-pooled.

3.5 Discussion

In this chapter, we consider a distributed environment that data are from multiple sources in the presence of missing data. Due to institutional policies or concerns about privacy, data are not allowed to be combined. It can also be shown from our simulation studies that instead of using our privacy-preserving methods with distributed analysis techniques, the standard MI method (MI-naïve) has a poor performance in terms of bias and MSE. Our numerical studies demonstrate that PPMI-H for the univariate missing data patterns and PPMICE for the general missing data patterns perform as well as the method using the pooled data. As we illustrate in Section 3.2.3, PPMICE algorithm is an iterative procedure and the update of individual variables depends on other variables through the imputation models. Therefor, it presents computational challenges since it requires a lot of communication. Moreover, the computation is hard to parallelize and is not ideal for distributed computing.

Table 3.1: Simulation results for estimating $\theta_0 = \theta_1 = \theta_2 = 1$ where a continuous variable has missing values. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; PPIPW-H, privacy-preserving IPW; MI-naïve, locally applying MI; MI-pooled, MI using pooled data; PPMI-H, privacy-preserving MI.

		$\hat{\theta}_0$					$\hat{\theta}_1$					$\hat{\theta}_2$				
n	Method	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)
200	GS	0.147	0.071	0.070	0.005	95.2	0.109	0.071	0.072	0.005	95.4	-0.759	0.142	0.143	0.020	94.7
	CC	-33.942	0.104	0.101	0.125	9.9	-9.962	0.094	0.094	0.019	80.9	-24.950	0.186	0.187	0.097	73.0
	IPW-pooled	-3.341	0.133	0.146	0.022	89.7	-2.091	0.124	0.149	0.022	90.2	-5.188	0.243	0.293	0.089	89.9
	PPIPW-H (K=5)	-3.341	0.133	0.146	0.022	89.0	-2.091	0.124	0.149	0.022	89.3	-5.188	0.243	0.293	0.089	89.5
	PPIPW-H (K=20)	-3.341	0.133	0.146	0.022	89.0	-2.091	0.124	0.149	0.022	89.3	-5.188	0.243	0.293	0.089	89.5
	MI-naïve (K=5)	0.137	0.099	0.095	0.009	95.9	-11.053	0.103	0.087	0.020	85.5	10.435	0.197	0.187	0.046	93.4
	MI-naïve (K=20)	-0.323	0.105	0.094	0.009	96.5	-80.156	0.131	0.105	0.654	0.1	79.235	0.225	0.196	0.666	6.6
	MI-pooled	0.318	0.095	0.096	0.009	95.3	-0.294	0.085	0.082	0.007	95.3	-0.018	0.181	0.182	0.033	94.8
	PPMI-H (K=5)	0.303	0.096	0.096	0.009	96.2	-0.483	0.086	0.082	0.007	95.6	0.146	0.178	0.182	0.033	94.1
	PPMI-H (K=20)	0.328	0.096	0.096	0.009	95.4	-0.449	0.086	0.082	0.007	95.5	0.171	0.178	0.182	0.033	94.1
1000	GS	0.010	0.032	0.031	0.001	96.3	-0.014	0.032	0.031	0.001	96.4	-0.186	0.063	0.061	0.004	96.2
	CC	-34.084	0.046	0.045	0.118	0.0	-10.198	0.041	0.043	0.012	31.3	-24.564	0.083	0.081	0.067	14.4
	IPW-pooled	-1.172	0.071	0.081	0.007	90.8	-1.382	0.067	0.088	0.008	87.0	-1.424	0.133	0.162	0.026	89.5
	PPIPW-H (K=5)	-1.172	0.071	0.081	0.007	91.0	-1.382	0.068	0.088	0.008	86.5	-1.424	0.133	0.162	0.026	89.1
	PPIPW-H (K=20)	-1.172	0.071	0.081	0.007	91.0	-1.382	0.068	0.088	0.008	86.5	-1.424	0.133	0.162	0.026	89.1
	MI-naïve (K=5)	0.065	0.043	0.042	0.002	95.2	-2.044	0.039	0.038	0.002	92.5	1.904	0.082	0.081	0.007	94.4
	MI-naïve (K=20)	0.105	0.044	0.042	0.002	97.1	-10.163	0.045	0.040	0.012	36.9	10.104	0.088	0.083	0.017	80.8
	MI-pooled	0.081	0.042	0.042	0.002	95.1	-0.127	0.037	0.037	0.001	94.6	-0.003	0.080	0.080	0.006	95.3
	PPMI-H (K=5)	0.065	0.043	0.042	0.002	95.8	-0.160	0.037	0.037	0.001	95.3	0.010	0.078	0.080	0.006	94.5
	PPMI-H (K=20)	0.070	0.043	0.042	0.002	95.4	-0.163	0.037	0.037	0.001	95.2	0.023	0.078	0.080	0.006	94.6

Table 3.2: Simulation results for estimating $\theta_0 = \theta_1 = \theta_2 = 1$ where a binary variable has missing values. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; PPIPW-H, privacy-preserving IPW; MI-naïve, locally applying MI; MI-pooled, MI using pooled data; PPMI-H, privacy-preserving MI.

		$\hat{\theta}_0$					$\hat{\theta}_1$					$\hat{\theta}_2$				
n	Method	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)
200	GS	0.205	0.102	0.103	0.011	95.4	-0.346	0.147	0.149	0.022	95.0	0.417	0.128	0.131	0.017	94.2
	CC	-33.361	0.130	0.133	0.129	27.1	-15.095	0.195	0.194	0.060	87.7	-28.896	0.183	0.188	0.119	62.6
	IPW-pooled	-2.541	0.165	0.184	0.035	90.4	-2.708	0.268	0.311	0.097	92.1	-6.506	0.236	0.295	0.091	85.4
	PPIPW-H (K=5)	-2.541	0.166	0.184	0.035	90.8	-2.708	0.267	0.311	0.097	92.3	-6.506	0.237	0.295	0.091	86.1
	PPIPW-H (K=20)	-2.589	0.165	0.187	0.035	90.1	-3.124	0.264	0.306	0.094	92.3	-6.713	0.234	0.295	0.091	85.9
	MI-naïve (K=5)	7.368	0.138	0.129	0.022	92.5	-12.330	0.210	0.191	0.052	93.1	5.125	0.147	0.145	0.024	94.0
	MI-naïve (K=20)	24.325	0.116	0.132	0.077	45.0	-40.232	0.191	0.209	0.205	46.9	14.990	0.138	0.140	0.042	80.1
	MI-pooled	0.479	0.140	0.142	0.020	95.1	-1.366	0.204	0.208	0.043	94.1	0.902	0.148	0.153	0.024	94.1
	PPMI-H (K=5)	0.475	0.140	0.141	0.020	95.1	-1.350	0.204	0.206	0.043	95.5	0.891	0.148	0.153	0.024	94.0
	PPMI-H (K=20)	0.488	0.139	0.143	0.021	93.8	-1.469	0.203	0.209	0.044	94.5	0.831	0.148	0.154	0.024	93.8
1000	GS	-0.273	0.046	0.046	0.002	95.1	0.226	0.066	0.066	0.004	93.9	0.035	0.057	0.058	0.003	94.5
	CC	-33.828	0.057	0.057	0.118	0.0	-14.363	0.087	0.084	0.028	62.8	-29.626	0.081	0.082	0.094	5.1
	IPW-pooled	-1.325	0.082	0.086	0.008	92.4	-0.165	0.139	0.150	0.022	93.4	-2.575	0.127	0.146	0.022	88.6
	PPIPW-H (K=5)	-1.325	0.082	0.086	0.008	92.1	-0.165	0.139	0.150	0.022	92.7	-2.575	0.127	0.146	0.022	88.5
	PPIPW-H (K=20)	-1.325	0.082	0.086	0.008	92.1	-0.165	0.139	0.150	0.022	92.7	-2.575	0.127	0.146	0.022	88.5
	MI-naïve (K=5)	1.407	0.063	0.062	0.004	94.9	-2.608	0.092	0.087	0.008	95.9	1.181	0.067	0.066	0.004	94.0
	MI-näive (K=20)	6.267	0.062	0.058	0.007	83.6	-10.415	0.093	0.083	0.018	83.5	4.395	0.066	0.064	0.006	90.0
	MI-pooled	-0.433	0.063	0.063	0.004	94.9	0.289	0.090	0.088	0.008	95.2	-0.079	0.066	0.067	0.004	94.5
	PPMI-H (K=5)	-0.398	0.063	0.063	0.004	94.3	0.236	0.090	0.088	0.008	95.7	-0.051	0.066	0.067	0.004	94.4
	PPMI-H (K=20)	-0.419	0.063	0.063	0.004	95.0	0.262	0.089	0.088	0.008	95.5	-0.071	0.066	0.067	0.004	95.0

Table 3.3: Simulation results for estimating $\theta_1 = \theta_2 = \theta_3 = 1$ where three continuous variables have missing values. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; MI-naïve, locally applying MI; MI-pooled, MI using pooled data; PPMICE, privacy-preserving MICE.

		$\hat{\theta}_1$					$\hat{\theta}_2$					$\hat{\theta}_3$				
n	Method	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)
	GS	-0.231	0.072	0.073	0.005	94.1	-0.166	0.072	0.071	0.005	96.0	-0.016	0.072	0.073	0.005	95.0
	CC	-13.441	0.234	0.252	0.082	87.9	-13.101	0.235	0.248	0.079	89.2	-11.031	0.237	0.250	0.075	90.0
	MI-naïve (K=5)	-39.782	0.159	0.116	0.172	25.3	-18.199	0.145	0.118	0.047	80.9	-21.536	0.148	0.115	0.059	74.1
200	MI-naïve (K=20)	-90.937	0.103	0.051	0.830	0.0	-85.916	0.132	0.074	0.744	0.0	-88.401	0.120	0.064	0.786	0.1
	MI-pooled	-3.791	0.137	0.129	0.018	95.1	-2.306	0.130	0.126	0.016	95.5	-1.962	0.132	0.127	0.017	96.0
	PPMICE (K=5)	-1.968	0.133	0.136	0.019	93.8	-1.604	0.128	0.131	0.017	94.5	-0.498	0.129	0.132	0.018	94.8
	PPMICE (K=20)	-1.968	0.133	0.136	0.019	93.8	-1.604	0.128	0.131	0.017	94.5	-0.498	0.129	0.132	0.018	94.8
	GS	-0.052	0.032	0.032	0.001	94.3	-0.014	0.032	0.030	0.001	95.5	-0.002	0.032	0.032	0.001	94.4
	CC	-11.786	0.095	0.096	0.023	75.8	-12.074	0.095	0.096	0.024	74.4	-11.868	0.095	0.098	0.024	75.6
	MI-naïve (K=5)	-10.507	0.064	0.053	0.014	64.9	-4.513	0.058	0.053	0.005	90.1	-5.868	0.059	0.053	0.006	86.8
1000	MI-naïve (K=20)	-38.264	0.073	0.052	0.149	0.0	-16.655	0.063	0.051	0.030	20.7	-20.140	0.065	0.051	0.043	8.4
	MI-pooled	-2.386	0.059	0.053	0.003	95.7	-1.181	0.056	0.054	0.003	95.2	-2.018	0.058	0.055	0.003	93.8
	PPMICE (K=5)	-0.097	0.056	0.056	0.003	95.3	-0.295	0.055	0.056	0.003	95.1	-0.145	0.056	0.057	0.003	93.9
	PPMICE (K=20)	-0.097	0.056	0.056	0.003	95.3	-0.295	0.055	0.056	0.003	95.1	-0.145	0.056	0.057	0.003	93.9

Table 3.4: Simulation results for estimating $\theta_0 = 0$, $\theta_1 = 1$, $\theta_2 = 1$ when the data are MNAR. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; MI-naïve, locally applying MI assuming MAR; MI-pooled, MI assuming MAR using pooled data; MI-pooled-RW, re-weighting the imputations from MI-pooled; PPMI-RW, privacy-preserving re-weighting the imputations from PPMI-H.

	$\hat{\theta}_0$					$\hat{\theta}_1$					$\hat{\theta}_2$				
	RBias(%)	SE	SD	MSE	CR (%)	RBias(%)	SE	SD	MSE	CR (%)	RBias(%)	SE	SD	MSE	CR (%)
GS	0.504	0.035	0.035	0.001	93.0	0.982	0.116	0.115	0.013	96.0	0.026	0.164	0.166	0.028	96.0
CC	32.835	0.041	0.041	0.109	0.0	-28.037	0.135	0.120	0.093	43.5	-15.051	0.187	0.180	0.055	86.5
MI-naïve	-1.593	0.036	0.036	0.002	92.5	-14.809	0.155	0.144	0.042	88.0	18.603	0.195	0.199	0.074	83.5
MI-pooled	-1.608	0.036	0.036	0.002	92.5	-14.487	0.156	0.145	0.042	88.5	18.338	0.195	0.201	0.074	84.5
MI-pooled-RW	-0.094	0.035	0.037	0.001	93.0	1.543	0.121	0.149	0.022	87.0	-0.852	0.168	0.208	0.043	88.5
PPMI-RW	-0.072	0.035	0.037	0.001	94.5	1.766	0.122	0.155	0.024	86.0	-1.139	0.169	0.210	0.044	88.0

Table 3.5: Regression coefficients estimates of the Georgia stroke registry data. NPO, nil per os, Latin for “nothing by mouth”, a medical instruction to withhold oral intake of food and fluids from a patient. P-value, (); 95% confidence interval, [ ].

Characteristics	CC	MI-naïve	MI-pooled	PPMICE
NIH stroke score	-1.95 (**<0.001**) [-2.70, -1.20]	-0.02 (0.951) [-0.69, 0.65]	-5.60 (**<0.001**) [-8.16, -3.04]	-5.67 (**<0.001**) [-8.34, -3.00]
EMS pre-notification	-3.17 (0.590) [-14.69, 8.35]	13.78 (0.302) [-12.69, 40.24]	-11.72 (0.439) [-42.02, 18.57]	-6.02 (0.697) [-37.05, 25.00]
Serum total lipid	-0.07 (0.201) [-0.18, 0.04]	-0.16 (0.205) [-0.41, 0.09]	-0.54 (**<0.001**) [-0.75, -0.33]	-0.53 (**<0.001**) [-0.74, -0.32]
Age	0.02 (0.936) [-0.51, 0.56]	-0.92 (**0.018**) [-1.67, -0.16]	-0.79 (**0.042**) [-1.54, -0.03]	-0.79 (**0.040**) [-1.55, -0.04]
Male (referent: female)	5.33 (0.372) [-6.37, 17.02]	14.40 (0.099) [-2.70, 31.50]	24.87 (**0.004**) [7.95, 41.80]	24.65 (**0.004**) [7.73, 41.56]
White (referent: African American)	-16.64 (**0.007**) [-28.82, -4.45]	-10.12 (0.263) [-27.85, 7.61]	-18.87 (**0.038**) [-36.69, -1.04]	-19.48 (**0.034**) [-37.51, -1.46]
Health insurance by medicare	-4.07 (0.617) [-20.04, 11.90]	-24.01 (**0.041**) [-47.05, -0.97]	-25.48 (**0.029**) [-48.29, -2.68]	-25.05 (**0.031**) [-47.88, -2.22]
Arrive in the daytime	4.94 (0.420) [-7.07, 16.96]	-12.67 (0.185) [-31.42, 6.09]	-25.87 (**0.005**) [-44.03, -7.72]	-24.25 (**0.010**) [-42.71, -5.79]
NPO	8.37 (0.393) [-10.84, 27.58]	62.45 (**<0.001**) [32.16, 92.74]	110.31 (**<0.001**) [64.20, 156.41]	112.61 (**<0.001**) [65.97, 159.24]
History of stroke	-2.57 (0.695) [-15.43, 10.29]	-29.98 (**0.007**) [-51.82, -8.14]	-33.13 (**0.003**) [-54.76, -11.50]	-32.42 (**0.003**) [-53.44, -11.40]
History of TIA	-16.30 (0.097) [-35.54, 2.94]	-62.56 (**0.001**) [-100.16, -24.96]	-87.63 (**<0.001**) [-120.66, -54.59]	-85.57 (**<0.001**) [-122.33, -48.81]
History of cardiac valve prosthesis	-27.25 (0.349) [-84.27, 29.78]	118.13 (**<0.001**) [58.77, 177.49]	140.91 (**<0.001**) [68.26, 213.56]	143.65 (**<0.001**) [71.78, 215.51]
Family history of stroke	-17.33 (0.406) [-58.18, 23.51]	-54.02 (0.075) [-113.58, 5.54]	-68.93 (0.065) [-142.22, 4.36]	-79.59 (0.059) [-162.30, 3.12]

Chapter 4

Privacy-Preserving Methods for Vertically Partitioned Incomplete Data

4.1 Introduction

Given the advantages of distributed data networks as described in Chapter 4, there is an ever-increasing need to develop statistical methods for analyzing data within such networks. The process of integrating data poses real privacy issues: data that are of restricted sensitivity may become highly sensitive after being integrated (pooled). For example, linking clinical diagnosis data with patients' demographic records leads to high sensitive data, and releasing them across networks (institutions) may suffer from high risk of disclosure. Research is rapidly progressing to propose novel privacy-preserving approaches that can overcome such privacy issues. Vaidya and Clifton (2004) adopt a conceptually simple definition of "privacy": a collaborating institution should learn nothing from any other institution's data. To protect privacy and reduce disclosure risk, it is common for institutions to manipulate (e.g. perturb or coarsen) the data prior to integrate (Willenborg and de Waal, 2012). However, low risk of disclosure caused by altering data, especially synthesizing data, brings reduced utilities and imprecise conclusions as well (Holan et al., 2010). An alternative approach for protecting against exposure of sensitive data is to compute and release only the summary statistics. Within the statistics literature, most attention has been drawn into the case of horizontally partitioned data. In Chapter 3, we investigate how to analyze incomplete horizontally partitioned data. In this chapter, we investigete a more common case that data are vertically partitioned.

Vertically partitioned data refer to the data from different institutions that have mutually exclusive characteristics for the same population. This is commonly present in real collaborations among different types of data providers. For instance, local and federal agencies, hospitals and private corporations with different information about the same population can work together to develop comprehensive quantitative models to produce meaningful results. A variety of privacy-preserving methods have been proposed to address statistical tasks including linear regression (Karr et al., 2009; Sanil et al., 2004) and logistic regression (Li et al., 2016). Other work on mining vertically partitioned data include linear discriminant analysis (Du et al., 2004), association rule mining (Vaidya and Clifton, 2002), support vector machine (Yu et al., 2006), naïve Bayes (Vaidya and Clifton, 2004) and k-means (Vaidya and Clifton, 2003). From a statistical perspective, some of these techniques proposed by computer scientists are incomplete in a way that, for example, coefficient estimates are provided, while standard errors and other essential statistics for inferences are ignored. Those neglect statistics are of decisive importance in some biomedical research, especially association studies. Karr et al. (2009) propose a protocol for model diagnostics via secure matrix multiplications. However, their method requires large communication costs and heavy computation when the number of institutions is large, and is thus not scalable.

Although the developments of privacy-preserving alternatives of the standard statistical learning techniques are extensive, research on how to deal with missing values for such vertically partitioned data is absent. In addition, with the prevalence of distributed networks and an increasing number of institutions participating, investigators experience missing values more frequently. These two factors motivate us to propose privacy-preserving methods for incomplete vertically partitioned data. Specifically, assuming the data follow a univariate missing data pattern, we propose two privacy-preserving approaches that couple distributed models (linear regression and logistic regression) with an inverse probability weighting (IPW) technique and a multiple imputations (MI) technique, respectively. Our privacy-preserving IPW for vertically partitioned data (PPIPW-V) first builds a distributed logistic regression model on the probability of observing a complete case, without disclosing individual-level data. Then we calculate the weights as the inverse of the estimated probabilities, and fit a weighted distributed linear regression model assuming our original analysis model of interest is a multiple linear regression. PPIPW-V can be easily extended to the case of logistic regression of a binary outcome variable on independent variables which are collected by different institutions. As introduced in Chapter 1, MI methods for handling missing data are popular and are shown to perform well in both literature and practice. We propose a privacy-preserving MI approach for vertically partitioned data (PPMI-V) assuming response variable is fully observed while one independent variable may be missing partially on a subset of records. Utilizing the technique of multiple imputation by chained equations (MICE), we can extend PPMI-V to be applicable to data that have general missing data patterns. We offer guidance and suggestions to calculate standard errors for both PPIPW-V and PPMI-V through bootstrap resampling.

The remainder of this chapter is organized as follows. In the beginning of Section 4.2, we formulate and describe settings and notation of missing data that are vertically partitioned. In Section 4.2.1 and 4.2.2, we formally develop PPIPW-V and PPMI-V. In Section 4.3 we show that the proposed methods for vertically partitioned data perform as well as using pooled data. We also provide some practical recommendations for applications. In Section 4.4, we generate synthetic incomplete data from the Georgia Coverdell Acute Stoke Registry (GCASR) data to mimic the case that data are vertically partitioned. PPIPW-V and PPMI-V as well as other methods are applied on the synthetic data for comparisons. This empirical study demonstrates the effectiveness of our proposed methods on real large samples. Section 4.5 concludes this chapter with discussions.

4.2 Methodology

The methods introduced in the previous chapter work only on horizontally partitioned data. In this section, we investigate another type of distributed data (i.e., vertically partitioned data) and propose privacy-preserving approaches that can address such data. In the same way as the previous "analysis model" setting, we consider the regression model $Y = \boldsymbol{X}\boldsymbol{\theta} + \epsilon$. The objective of the regression analysis is to estimate regression coefficients $\boldsymbol{\theta}$, when the covariate $\boldsymbol{X}$ is subject to missing values. For vertically partitioned data from a distributed environment with K institutions (refer as sites in the formula), we let $\boldsymbol{X} = (\boldsymbol{X}^{site_1}, ..., \boldsymbol{X}^{site_K})$, where $\boldsymbol{X}^{site_k}$ is a set of covariates collected from the k-th institution. Such scenario assumes that single institution has exclusive variables of the same population. We assume that the outcome variable Y is accessible to all institutions. Unlike the assumption of $p = 2$ in Section 3.2, we investigate any p and denote the number of covariates in the k-th institution by p_k and $\sum_{k=1}^{K} p_k = p$. Without loss of generality, we assume $\boldsymbol{X}^{site_1}$ has a vector of all 1's to include an "intercept" term in the regression model, i.e., $\boldsymbol{X}^{site_1} = (\mathbb{1}, X_1, ..., X_{p_1})$. We consider a univariate missing data pattern where X_1 (from $\boldsymbol{X}^{site_1}$) has missing values.

4.2.1 Privacy-preserving inverse probability weighting for vertically partitioned data

To apply IPW on vertically partitioned data, we need to develop a distributed logistic regression model for the weights and a distributed linear regression model for the weighted subjects.

Denote the predictors of the probability of observing X_1 by $\boldsymbol{Z} = (\mathbb{1}, Y, X_2, ..., X_p)$. For notational convenience, let $\boldsymbol{Z} = (\boldsymbol{Z}^{site_1}, ..., \boldsymbol{Z}^{site_K})$, where $\boldsymbol{Z}^{site_1} = (\mathbb{1}, Y, X_2, ..., X_{p_1})$ and $\boldsymbol{Z}^{site_k} \equiv \boldsymbol{X}^{site_k}$, for $k \geq 2$.

The distributed Newton Raphson algorithm for logistic regression given in (3.3) does not apply here because $\boldsymbol{Z}^T \boldsymbol{W}^{old} \boldsymbol{Z} \neq \sum_{k=1}^{K} (\boldsymbol{Z}^{site_k})^T \boldsymbol{W}^{old} \boldsymbol{Z}^{site_k}$. Jaakkola and Haussler (1999) introduce an alternative approach to optimize the logistic regression model by dual optimization. The original maximization of the log-likelihood of a primal problem is replaced with the minimization of the dual form log-likelihood, which guarantees the same optimum. The linear decomposition becomes feasible for the dual optimization. In the following, we use a new response indicator s_i, taking value 1 if individual i is fully observed and -1 otherwise, to better represent the primal form of the log-likelihood function.

The logistic regression model for the response indicator becomes $Pr(s_i = \pm 1 | \boldsymbol{z}) = 1/(1 + \exp(-s_i \boldsymbol{z}_i^T \boldsymbol{\beta}))$, where $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}$ is a vector of nuisance parameters to be estimated and $\boldsymbol{z}_i$ is the i-th row of $\boldsymbol{Z}$. The primal problem is to maximize the log-likelihood $l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log(1 + \exp(-s_i \boldsymbol{z}_i^T \boldsymbol{\beta})) - \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}/2$. The penalty $\lambda \boldsymbol{\beta}^T \boldsymbol{\beta}/2$ is introduced to give a superior generalization performance, especially when p is large. Instead of solving the primal problem, we solve the dual problem which is represented by dual parameter $\boldsymbol{\psi} \in \mathbb{R}^{n\times 1}$ as:

$$\min_{\boldsymbol{\psi}} J(\boldsymbol{\psi}) = \frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{i'=1}^{n}\psi_i\psi_{i'}s_i s_{i'}\boldsymbol{z}_i^T\boldsymbol{z}_{i'} - \sum_{i=1}^{n}H(\psi_i), \tag{4.1}$$

where $H(\psi_i) = -\psi\log\psi - (1-\psi)\log(1-\psi)$. It is easy to see that the linear kernel $\boldsymbol{z}_i^T\boldsymbol{z}_{i'}$ in Equation (4.1) can be linearly decomposed by institutions as $\boldsymbol{z}_i^T\boldsymbol{z}_{i'} = \sum_k(\boldsymbol{z}_i^{site_k})^T\boldsymbol{z}_{i'}^{site_k}$. Such decomposition builds the foundation of a privacy-preserving distributed logistic regression model over vertically partitioned data. That is, each institution computes the dot products $(\boldsymbol{z}_i^{site_k})^T\boldsymbol{z}_{i'}^{site_k}$ and shares them to calculate $\boldsymbol{z}_i^T\boldsymbol{z}_{i'}$ of each pair of individuals. Since the dot product is a scalar, the exposure of it does not lead to the disclosure of $\boldsymbol{z}_i$. Newton-Raphson algorithm is applied to optimize $\boldsymbol{\psi}$ via iterative procedures until convergence: $\hat{\boldsymbol{\psi}}^{new} = \hat{\boldsymbol{\psi}}^{old} - J''(\hat{\boldsymbol{\psi}}^{old})^{-1}J'(\hat{\boldsymbol{\psi}}^{old})$. With the estimated dual parameters $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, ..., \hat{\psi}_n)^T$, we can get the estimated primal parameters $\hat{\boldsymbol{\beta}} = ((\hat{\boldsymbol{\beta}}^{site_1})^T, ..., (\hat{\boldsymbol{\beta}}^{site_K})^T)^T$ by sending $\hat{\boldsymbol{\psi}}$ to each institution: $\hat{\boldsymbol{\beta}}^{site_k} = \lambda^{-1}\sum_{i=1}^{n}\hat{\psi}_i s_i \boldsymbol{z}_i^{site_k}$. Then, we can obtain the weight for individual i by $\hat{w}_i = 1/\hat{p}_i = 1 + \exp(-\sum_{k=1}^{K}(\boldsymbol{z}_i^{site_k})^T\hat{\boldsymbol{\beta}}^{site_k})$. Note that the dot product $(\boldsymbol{z}_i^{site_k})^T\hat{\boldsymbol{\beta}}^{site_k}$ is calculated locally by each institution and then shared to others.

We then establish a weighted distributed linear regression model in this vertically partitioned setting. The objective function can be written in the matrix form: $F(\boldsymbol{\theta}) = (Y - \boldsymbol{X\theta})^T\boldsymbol{V}(Y - \boldsymbol{X\theta})$, where $\boldsymbol{V}$ is the diagonal matrix of weights. In the case of IPW, $\boldsymbol{V} = diag(\{r_i\hat{w}_i\}_{i=1}^n)$. The quadratic programming problem of minimizing $F(\boldsymbol{\theta})$ can be solved by the following derivative-free modified Powell's algorithm proposed by Sanil et al. (2004):

- *Initialization*: Select an arbitrary orthogonal basis for $\mathbb{R}^{(p+1)}$: $\boldsymbol{d}^{(1)}, ..., \boldsymbol{d}^{(p+1)}$. Pick an arbitrary starting point $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{(p+1)}$

- *Iteration*: Repeat the following steps $p+1$ times.

    - Set $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$

```
– For $j = 1, 2, ..., p, p+1$:
    ∗ Let $\delta = \arg\min_\delta F(\boldsymbol{\theta} + \delta \boldsymbol{d}^{(j)})$
    ∗ Set $\boldsymbol{\theta} = \boldsymbol{\theta} + \delta \boldsymbol{d}^{(j)}$
– For $j = 1, 2, ..., p$: Set $\boldsymbol{d}^{(j)} = \boldsymbol{d}^{(j+1)}$
– Set $\boldsymbol{d}^{(p+1)} = \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}$
    ∗ Let $\delta = \arg\min_\delta F(\boldsymbol{\theta} + \delta \boldsymbol{d}^{(p+1)})$
    ∗ Set $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \delta \boldsymbol{d}^{(p+1)}$
```

Of note, for the objective function of sums of weighted errors, given any direction $\boldsymbol{d}$,

$$\delta = \arg\min_\delta F(\boldsymbol{\theta} + \delta \boldsymbol{d}) = \frac{(Y - \boldsymbol{X\theta})^T \boldsymbol{V}(\boldsymbol{Xd})}{(\boldsymbol{Xd})^T \boldsymbol{V}(\boldsymbol{Xd})} = \frac{\boldsymbol{\gamma}^T \boldsymbol{V\eta}}{\boldsymbol{\eta}^T \boldsymbol{V\eta}},$$

where $\boldsymbol{\gamma} = Y - \boldsymbol{X\theta}$ and $\boldsymbol{\eta} = \boldsymbol{Xd}$. Similar to the data are vertically partitioned (i.e., $\boldsymbol{X} = (\boldsymbol{X}^{site_1}, ..., \boldsymbol{X}^{site_K})$), we partition the direction $\boldsymbol{d}$ and the vector of parameters $\boldsymbol{\theta}$ as $\boldsymbol{d} = ((\boldsymbol{d}^{site_1})^T, ..., (\boldsymbol{d}^{site_K})^T)^T$ and $\boldsymbol{\theta} = ((\boldsymbol{\theta}^{site_1})^T, ..., (\boldsymbol{\theta}^{site_K})^T)^T$ accordingly. Therefore, $\boldsymbol{\gamma} = Y - \sum_{k=1}^K \boldsymbol{X}^{site_k} \boldsymbol{\theta}^{site_k}$ and $\boldsymbol{\eta} = \sum_{k=1}^K \boldsymbol{X}^{site_k} \boldsymbol{d}^{site_k}$. Such linear decompositions allow us to obtain δ by only sharing the locally calculated summary statistics (i.e., $\boldsymbol{X}^{site_k} \boldsymbol{\theta}^{site_k}$, $\boldsymbol{X}^{site_k} \boldsymbol{d}^{site_k}$). Powell (1964) showed that if $F(\boldsymbol{\theta})$ is a quadratic function, the above algorithm would yield the exact minimizer $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$. The asymptotic distribution of the weighted parametric estimate of $\boldsymbol{\theta}$ is derived by Wang et al. (1997). Their rigorous estimated covariance matrix is sophisticated and difficult to calculate in the distributed environments. We use bootstrap to approximate the standard error of $\hat{\boldsymbol{\theta}}$, for practical considerations. Generating bootstrap data $(Y, \boldsymbol{X})^b$ repeatedly using vertically distributed data is intuitive, we sample the indices $(1, ..., n)$ with replacement and share them to all institutions, which prepare (arrange) the dataset accordingly.

4.2.2 Privacy-preserving multiple imputations for vertically partitioned data

Similar to Section 3.2.2, we propose a distributed multiple imputation method for vertically partitioned data assuming MAR. Suppose the missing variable X_1 is continuous and follows a normal distribution given Y and other covariates. That is, $X_1 = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim_{iid} N(0, \sigma^2)$. We can use complete cases to estimate parameter $\boldsymbol{\alpha}$. The objective function (sum of squared residuals) is $F(\boldsymbol{\alpha}) = (X_1 - \boldsymbol{Z}\boldsymbol{\alpha})^T\boldsymbol{V}(X_1 - \boldsymbol{Z}\boldsymbol{\alpha})$, with $\boldsymbol{V} = diag(\{r_i\}_{i=1}^n)$. Since $F(\boldsymbol{\alpha})$ is in a quadratic form, derivative-free modified Powell's algorithm can be directly applied to get the least square estimator $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$. As we illustrate in Section 4.2.1, the intermediate value δ can calculate using summary statistics from each institution, leading to obtaining $\hat{\boldsymbol{\alpha}}$ with confidentiality. We can then take advantage of the bootstrap technique to estimate the variance of $\hat{\boldsymbol{\alpha}}$, denoted by $\hat{\boldsymbol{V}_{\boldsymbol{\alpha}}}$. The multiple imputation method first draws a value ($\hat{\boldsymbol{\alpha}}^{(m)}$, $\hat{\sigma}^{(m)2}$) from the posterior distribution of $(\boldsymbol{\alpha}, \sigma^2)$, where $\hat{\boldsymbol{\alpha}}^{(m)}$ is drawn from a multivariate normal distribution with mean $\hat{\boldsymbol{\alpha}}$ and variance matrix $\hat{\boldsymbol{V}_{\boldsymbol{\alpha}}}$, and $\hat{\sigma}^{(m)2}$ is drawn from $\sum_{i=1}^n r_i(x_{1,i} - \sum_{k=1}^K (\boldsymbol{z}_i^{site_k})^T \hat{\boldsymbol{\alpha}}^{(m)site_k})^2 / \chi^2_{(\sum r_i - p - 1)}$. For individuals i missing X_1, given observed $\boldsymbol{z}_i$, $X_{1,i}$ is then drawn from a $N(\boldsymbol{z}_i^T \hat{\boldsymbol{\alpha}}^{(m)}, \hat{\sigma}^{(m)2})$, where $\boldsymbol{z}_i^T \hat{\boldsymbol{\alpha}}^{(m)} = \sum_{k=1}^K (\boldsymbol{z}_i^{site_k})^T \hat{\boldsymbol{\alpha}}^{(m)site_k}$ is linearly decomposable. We repeat the above procedure M times and create M multiply imputed datasets. Each dataset is then processed by the aforementioned Modified Powell's algorithm, resulting in M estimates $\{\hat{\boldsymbol{\theta}}^{(m)}\}_{m=1}^M$ of the parameters of interest $\boldsymbol{\theta}$. The final estimate can be obtained by combining $\{\hat{\boldsymbol{\theta}}^{(m)}\}_{m=1}^M$ using Rubin's rules.

4.3 Simulation Studies

We examine in this section the performance of the proposed methods. Suppose our vertically partitioned data $\boldsymbol{X}$ consist $p = 6$ independent variables from $K = 3$ institutions. We assume each institution possesses two independent variables and has access to the outcome variable Y. We are interested in a multiple linear regression:

$$\begin{aligned} Y &= \boldsymbol{X\theta} + \epsilon \\ &= (\overbrace{\mathbb{1}, X_1, X_2}^{site_1}, \overbrace{X_3, X_4}^{site_2}, \overbrace{X_5, X_6}^{site_3}) \times (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)^T + \epsilon, \end{aligned}$$

when X_1 has missing values.

We generate $X_2, ..., X_6$ independently from the uniform distribution on (-1, 1). We then generate X_1 from a normal distribution with mean $\sum_{j=2}^{5} X_j/\sqrt{5}$ and variance 1. The continuous variable Y is generated from $\mathcal{N}(\boldsymbol{X\theta}, \sigma^2)$, where $\boldsymbol{\theta} = (1, 1, 0, 1, 0, 1, 0)^T$ and $\sigma^2 = 1$. We consider two scenarios for the selection probabilities as follows:

$$\begin{aligned} &\text{Scenario 1.} \qquad Pr(R = 1|Y, \boldsymbol{X}) = \{1 + \exp(-1.6 + Y + X_3 + X_5)\}^{-1} \\ &\text{Scenario 2.} \qquad Pr(R = 1|Y, \boldsymbol{X}) = \{1 + \exp(3 + 2Y + 2X_3 + 2X_5)\}^{-1}. \end{aligned}$$

Since the missingness of X_1 does not depend on itself, the data are missing at random. Figure 4.1 presents examples of the frequency plots of the probabilities in the two scenarios using $n = 1000$ subjects. While both scenarios lead to about the same 42% of subjects missing X_1, Scenario 1 provides relatively stable weights compared to Scenario 2, where weights are defined as the inverse of the probabilities.

We compare analyses based on standard approaches, inverse probability weighting methods and multiple imputation. Specifically, we consider the following seven methods:

- Gold standard (GS): the analysis on n subjects with underlying true values for missing data

- Complete-cases (CC): the analysis on fully observed subjects using a distributed linear regression

- IPW-pooled: the standard IPW approach on pooled data throughout the whole process of estimating weights and fitting weighted linear regression

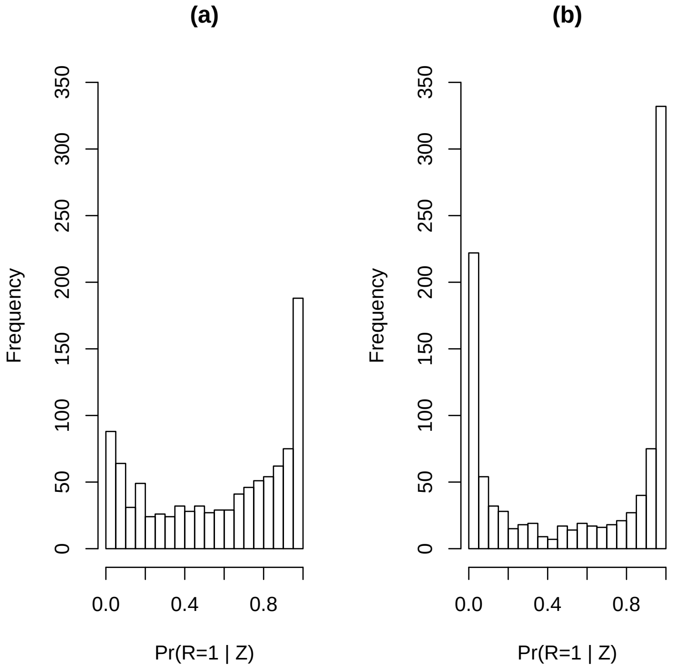


Figure 4.1: Histograms of the probabilities of observing a complete case in (a) Scenario 1 and (b) Scenario 2. using 1000 subjects

- PPIPW-V: the proposed privacy-preserving IPW for vertically partitioned data; see Section 4.2.1

- MI-naive: each institution imputes the missing data using their own data, following by applying a distributed linear regression for the analysis model

- MI-pooled: the standard MI approach on pooled data throughout the whole process of predicting missing values (M=100 times) and fitting a standard linear regression.

- PPMI-V: the proposed privacy-preserving MI with M=100 imputed datasets for vertically partitioned incomplete data; see Section 4.2.2

Table 4.1 shows the results of Scenario 1. The data are missing at random and the missingness mechanism depends on the outcome variable Y. We first illustrate the performances of standard IPW and MI using pooled data (i.e., IPW-pooled and MI-pooled). IPW-pooled generates unbiased estimators while CC provides a large bias. This finding is more clear when $n = 1000$, that the relative bias of IPW-pooled is negligible. This means that by applying the weights, IPW-pooled can, as expected, reduce the bias compared to CC which also uses the complete cases only. However, the results show that the estimators of IPW-pooled, even though unbiased, are still under covered by 95% confidence intervals. Of note, IPW-pooled also possesses a much larger SE than others. MI-pooled method, on the contrary, yields an unbiased estimator with a relatively small SE. It also has good coverages that are all close to the 95% level. In terms of vertically partitioned incomplete data, a naïve way of handling missing values is to replace them with predicted ones using other covariates from the same institution. We denote this method by MI-naïve. The imputation models in MI-naïve are improper since the missing variable may actually depend on covariates of other institutions. Thus, ignoring those covariates will generally lead to biased estimators. As shown in Table 4.1, MI-naïve has large biases and serious low coverage rates, which correspond to the fact the it is inconsistent.

PPIPW-V and PPMI-V inherit the property of standard IPW and MI using pooled data, respectively. They perform in the similar way to their corresponding versions of non distributed methods. In other words, PPIPW-V gives unbiased results but large SEs, as IPW-pool; PPMI-V provides unbiased estimators with small SEs, as MI-pooled. This interesting finding confirms that our proposed privacy-preserving methods on vertically partitioned data work as well as the methods used on pooled data. They offer solutions to the missing data problem in distributed data networks, by providing meaningful results without individual-level data.

Table 4.2 displays the results of Scenario 2, which has unstable weights. A considerable number of probabilities of observing a complete case are very close to 0, leading to extremely large weights. It is possible that in practice, the logistic model of missingness yields very large weights for some individuals with moderate weights, due to lack of fit. On one hand, the results show that both IPW-pooled and PPIPW-V have biased estimators that give rise to large MSE. MI-pooled and PPMI-V, on the other hand, do not require the specification of missingness model and perform well in this scenario.

Table 4.1: Simulation results for estimating $\theta_1 = \theta_3 = \theta_5 = 1$ based on 1000 Monte Carlo replications for Scenario 1 with sample size $n = 200$ or 1000. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval

		$\hat{\theta}_1$					$\hat{\theta}_3$					$\hat{\theta}_5$				
n	Method	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)
200	GS	0.058	0.065	0.066	0.004	95.1	0.259	0.127	0.131	0.017	93.2	-0.260	0.127	0.127	0.016	96.1
	CC	-9.427	0.088	0.087	0.016	80.5	-19.702	0.169	0.174	0.069	77.9	-20.409	0.170	0.163	0.068	76.9
	IPW-pooled	-2.964	0.113	0.139	0.020	88.3	-3.832	0.220	0.264	0.071	88.6	-6.445	0.220	0.265	0.074	88.1
	PPIPW-V	-3.293	0.108	0.129	0.018	89.0	-6.454	0.209	0.247	0.065	88.5	-8.787	0.210	0.243	0.067	87.1
	MI-naive	-3.227	0.083	0.080	0.007	94.3	-9.313	0.158	0.139	0.028	93.2	-9.740	0.158	0.140	0.029	93.7
	MI-pooled	-1.897	0.079	0.078	0.006	95.0	0.847	0.161	0.156	0.024	95.8	0.583	0.161	0.159	0.025	95.5
	PPMI-V	-1.620	0.074	0.079	0.006	92.9	1.410	0.167	0.158	0.025	96.0	1.265	0.166	0.160	0.026	95.5
1000	GS	-0.096	0.029	0.028	0.001	95.0	0.128	0.056	0.056	0.003	94.8	-0.117	0.056	0.056	0.003	94.9
	CC	-9.451	0.039	0.039	0.010	31.1	-19.975	0.075	0.074	0.045	23.9	-20.052	0.075	0.076	0.046	25.0
	IPW-pooled	-0.447	0.065	0.078	0.006	89.6	-1.318	0.122	0.142	0.020	89.6	-1.740	0.122	0.142	0.021	89.8
	PPIPW-V	-0.467	0.064	0.077	0.006	89.6	-1.867	0.121	0.140	0.020	88.0	-2.252	0.121	0.141	0.020	89.0
	MI-naive	-2.419	0.035	0.034	0.002	90.9	-9.900	0.069	0.061	0.013	72.2	-9.994	0.069	0.063	0.014	71.4
	MI-pooled	-0.257	0.034	0.033	0.001	95.4	0.392	0.070	0.069	0.005	95.4	0.135	0.071	0.071	0.005	94.5
	PPMI-V	-0.209	0.032	0.033	0.001	93.8	0.126	0.072	0.069	0.005	96.1	-0.086	0.072	0.070	0.005	95.7

Table 4.2: Simulation results for estimating $\theta_1 = \theta_3 = \theta_5 = 1$ based on 1000 Monte Carlo replications for Scenario 2 with sample size $n = 200$ or 1000. RBias, mean relative bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval

		$\hat{\theta}_1$					$\hat{\theta}_3$					$\hat{\theta}_5$				
n	Method	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)	RBias (%)	SE	SD	MSE	CR (%)
200	GS	-0.633	0.065	0.065	0.004	95.5	-1.422	0.126	0.137	0.019	90.5	-0.340	0.126	0.117	0.014	98.5
	CC	-15.239	0.087	0.082	0.030	57.5	-31.332	0.169	0.175	0.128	55.5	-30.333	0.168	0.166	0.120	55.0
	IPW-pooled	-9.441	0.110	0.122	0.024	81.5	-16.927	0.227	0.312	0.125	76.0	-15.376	0.224	0.273	0.098	78.0
	PPIPW-V	-10.797	0.098	0.096	0.021	76.5	-24.179	0.198	0.236	0.114	71.0	-22.149	0.197	0.203	0.090	73.0
	MI-naive	-4.405	0.081	0.075	0.007	93.5	-12.129	0.161	0.142	0.035	90.5	-11.445	0.161	0.149	0.035	88.0
	MI-pooled	-3.013	0.079	0.076	0.007	94.5	-0.078	0.167	0.168	0.028	93.5	1.295	0.165	0.168	0.028	95.5
	PPMI-V	-2.657	0.074	0.076	0.006	94.5	0.633	0.170	0.167	0.028	93.5	1.986	0.168	0.168	0.029	96.5
1000	GS	0.098	0.029	0.028	0.001	94.2	0.553	0.056	0.058	0.003	93.2	-0.907	0.057	0.053	0.003	95.3
	CC	-14.640	0.039	0.041	0.023	4.2	-31.422	0.075	0.076	0.104	1.1	-33.005	0.075	0.075	0.114	0.0
	IPW-pooled	-5.019	0.070	0.096	0.012	76.3	-7.326	0.150	0.254	0.070	71.6	-11.577	0.151	0.210	0.057	72.6
	PPIPW-V	-5.998	0.062	0.081	0.010	70.5	-11.713	0.132	0.210	0.058	64.2	-15.428	0.131	0.172	0.053	61.1
	MI-naive	-2.446	0.035	0.035	0.002	88.9	-11.166	0.071	0.063	0.016	66.3	-12.346	0.071	0.062	0.019	61.6
	MI-pooled	-0.307	0.034	0.034	0.001	95.8	1.365	0.072	0.074	0.006	93.7	-0.189	0.073	0.073	0.005	95.3
	PPMI-V	-0.170	0.033	0.034	0.001	93.7	1.216	0.074	0.074	0.006	95.8	-0.322	0.074	0.074	0.005	95.8

4.4 Data Example

In this section, we conduct an empirical study using real data to evaluate the effectiveness of our approaches. The data are collected and pooled by the Georgia Coverdell Acute Stoke Registry (GCASR). To simplify the regression setting, we are only interested in the effect of four characteristics (i.e., Gender, Race, NIH stroke score, History of stroke) on arrival-to-CT time. We assume that the pooled data are actually from two institutions, where the first institution has patients' demographic information (e.g., gender and race) and the second institution has clinical information (e.g., NIH stroke score and history of stroke). The outcome variable arrival-to-CT time is accessible to both institutions. We first select 31,918 patients with observations on all four independent variables and the dependent variable. The analysis on this dataset is considered as gold standard (GS). Next, we generate the missing data by artificially assigning some patients to be missing NIH stroke score through the model

$$Pr(X_3 \text{ is missing}) = \{1 + \exp(5 - Y - X_1 - X_2 - X_4)\}^{-1},$$

where $Y = \log(\text{arrival-to-CT time})$, $X_1 = 1$ if male and 0 otherwise, $X_2 = 1$ if White and 0 otherwise, $X_3 =$ NIH stroke score, $X_4 = 1$ if the patient has a history of stroke and 0 otherwise. About 45% of patients are missing NIH stroke score according to the above criterion.

Table 4.3 presents estimates, SEs and p-values of the analyses of applying seven methods noted in Section 4.3. We consider $M = 20$ and $M = 100$ imputations for MI methods (MI-pooled and PPMI-V). The results are quite close so we only present those of $M = 20$. Once again, the facts that PPIPW-V behaves in the same way as IPW-pooled and PPMI-V performs as well as MI-pooled, are confirmed by Table 4.3. Based on our analysis, it appears that there is a significant negative association between arrival-to-CT time and NIH stroke score by any of the estimates. The same finding is observed between arrival-to-CT time and race. The result from the CC analysis shows a negative effect of gender on arrival-to-CT time, while this conclusion does not hold by other methods. History of stroke is not shown to be statistically significant by IPW-pooled and PPIPW-V. In terms of the values of the estimates, MI-pooled and PPMI-V are closest to GS in general. MI-pooled and PPMI-V also provide relatively smaller SEs than other methods. The estimate of NIH stroke score using MI-naïve is only half of that using GS.

Table 4.3: Regression coefficients estimates of the Georgia stroke registry data.

Characteristics	Methods	Estimate	SE	P-value
Male (referent: female)	GS	0.073	0.014	<0.001
	CC	-0.155	0.018	<0.001
	IPW-pooled	0.106	0.031	<0.001
	PPIPW-V	0.106	0.031	<0.001
	MI-nave	0.050	0.014	<0.001
	MI-pooled	0.069	0.014	<0.001
	PPMI-V	0.068	0.014	<0.001
White (referent: African American)	GS	-0.159	0.014	<0.001
	CC	-0.353	0.018	<0.001
	IPW-pooled	-0.213	0.031	<0.001
	PPIPW-V	-0.213	0.031	<0.001
	MI-nave	-0.138	0.014	<0.001
	MI-pooled	-0.155	0.014	<0.001
	PPMI-V	-0.154	0.014	<0.001
NIH stroke score	GS	-0.032	0.001	<0.001
	CC	-0.024	0.001	<0.001
	IPW-pooled	-0.034	0.002	<0.001
	PPIPW-V	-0.034	0.002	<0.001
	MI-nave	-0.014	0.001	<0.001
	MI-pooled	-0.028	0.001	<0.001
	PPMI-V	-0.027	0.002	<0.001
History of stroke	GS	-0.041	0.016	0.012
	CC	-0.227	0.019	<0.001
	IPW-pooled	-0.033	0.028	0.122
	PPIPW-V	-0.033	0.028	0.122
	MI-nave	-0.052	0.017	0.001
	MI-pooled	-0.037	0.016	0.024
	PPMI-V	-0.037	0.016	0.024

4.5 Discussion

The privacy-preserving approaches that we present provide promising results for handling vertically partitioned incomplete data. In specific, PPIPW-V models the weights through a logistic regression and solves it's corresponding dual problem that utilizes summary statistics only. Then, we weight complete cases based on the estimated weights and solve an objective function of quadratic form by a derivative-free modified Powell's algorithm.

The calculations within the algorithm can be linearly partitioned among institutions. The final least-squares estimate is proved to minimize the objective function. Our numeric studies shows that PPIPW-V performs in the same way as IPW-pooled. Based on the issue proposed in many IPW research, we should pay more attentions on unstable weights, which would deteriorate the utilizations of IPW. PPIPW-V also tends to produce larger standard errors than other methods on missing data. Another privacy-preserving method that we propose is PPMI-V, which has the same superior performance as MI-pooled. PPMI-V is flexible and can be extended to general missing data patterns and the case that a binary variable is subject to missing values. In addition, for both PPIPW-V and PPMI-V, we provide a privacy-preserving way to calculate standard errors through bootstrap resampling.

Chapter 5

Summary and Future Work

5.1 Summary

This dissertation develops and investigates methods on incomplete big data. One major challenge of such large volume datasets is how to handle their high dimensions. The first project seeks to develop multiple imputation (MI) methods for general missing data patterns in the presence of high-dimensional data. The proposed methods couple MI with regularized regression, which is widely used for model trimming. Following the framework of "multivariate imputation by chained equations" (MICE), the methods can handle the case that more than one variable is subject to missing values.

The second and third projects develop statistical methods for handling missing data in distributed data networks where data are horizontally or vertically partitioned. The challenge here is how to address missing data problem without pooling the data. In Chapter 3, we propose privacy-preserving IPW (PPIPW-H) and MI (PPMI-H) for horizontally partitioned data assuming MAR. PPIPW-H utilizes a distributed logistic regression along with a distributed weighted linear regression. We further propose sensitivity analysis under MNAR and present a modified privacy-preserving MI by re-weighting (PPMI-RW). The effectiveness of our approaches is demonstrated through extensive simulation studies. In Chapter 4, we develop two privacy-preserving methods that address vertically partitioned incomplete data that are MAR.

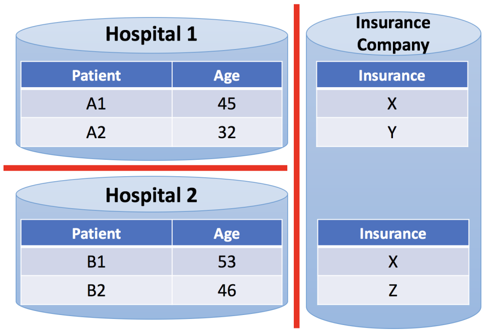
5.2 Future Work

In addition to the aforementioned work that we have done, we expect future works focusing on the following directions.

First, we are interested in the robustness of inferences of multiple imputations when the missing data are MNAR. As we know, multiple imputation method is widely used in practice, but with untestable assumption that data are MAR. All three of my topics utilize and implement MI. The first topic uses MICE whose joint distribution may not be consistent with univariate imputation models. Such situation may become even worse for data that are MNAR. The second topic studies sensitivity analysis for MNAR assumption through re-weighting MI. The analysis, instead, has a strong assumption for the response model, with a sensitivity parameter that needs to be pre-specified. The third topic proposes a privacy-preserving MI for vertically partitioned data. The method still assumes that data are MAR. Therefore, sensitivity analysis should be conducted for the MI methods on each topic.

Second, the doubly robust methods for inverse probability weighting on partitioned data can be investigated. Standard IPW provides consistent estimators when the weights are correctly estimated. However, it is inefficient because most of the information from the incomplete cases is not used. Robins et al. (1994) propose augmented IPW (AIPW) method which achieves double robustness, meaning that the estimators are consistent as long as either the missingness model or the regression model is correctly specified. It is of interest to bring AIPW method into distributed environments and propose privacy-preserving AIPW.

Third, more research can be conducted on complex partitioned data. In Chapter 3 and 4, we investigate data that are either horizontally partitioned or vertically partitioned. It is also common to encounter complex partitions of data in distributed data networks. Figure 5.1 shows an example of such type of partitioned data. There is no existing work on analyzing such complex partitions of data with or without missing values.



Patient	Age
A1	45
A2	32

Insurance
X
Y

Patient	Age
B1	53
B2	46

Insurance
X
Z

Figure 5.1: An example of complex partitions of data from 3 sites

Appendix A

Appendix for Chapter 2

A.1 Details of MICE-DURR for three types of data

We start the iterative procedure with some initial values. For example, all the elements in $Z_{mis,j}$ are filled in with the average of the observed values of Z_j $(j = 1, 2, ..., l)$. Define the corresponding initial completed dataset as $Z^{(0)}$.

In the m-th iteration:

(i) If Z_j follows a Gaussian distribution, the model is

$$Z^*_{j,obs} = \theta_{0,j}\mathbf{1}_{r^*_j} + \mathbf{W}^{*(m)}_{j,obs}\boldsymbol{\theta}_j + \boldsymbol{\epsilon}_j, \tag{A.1}$$

where r^*_j is the number of cases with observed Z^*_j and $\boldsymbol{\epsilon}_j \sim N(0, \sigma^2_j\mathbf{I}_{r^*_j})$.

A regularized regression method is used to fit model (A.1). The parameter estimates can be obtained as follows:

$$(\widehat{\theta}^{(m)}_{0,j}, \widehat{\boldsymbol{\theta}}^{(m)}_j) = \underset{(\theta_{0,j},\boldsymbol{\theta}_j)}{\text{argmin}}[-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; Z^*_{j,obs}, \mathbf{W}^{*(m)}_{j,obs}) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function. We consider the mean of squared residuals as an estimate of σ^2_j, denoted by $\widehat{\sigma}^{2(m)}_j$.

$Z_{j,mis}$ is predicted with $Z^{(m)}_{j,mis}$ by drawing randomly from the predictive distribution $N(\widehat{\theta}^{(m)}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}^{(m)}_{j,mis}\widehat{\boldsymbol{\theta}}^{(m)}_j, \widehat{\sigma}^{2(m)}_j\mathbf{I}_{n-r_j})$. Let $Z^{(m)}_j = (Z^{(m)}_{j,mis}, Z_{j,obs})$.

(ii) If Z_j follows a Bernoulli distribution, the model is

$$logit(Z^*_{j,obs} = 1|\mathbf{W}^{*(m)}_{j,obs}) = \theta_{0,j}\mathbf{1}_{r^*_j} + \mathbf{W}^{*(m)}_{j,obs}\boldsymbol{\theta}_j, \tag{A.2}$$

A regularized regression method is used to fit model (A.2). The parameter estimates can be obtained as follows:

$$(\widehat{\theta}^{(m)}_{0,j}, \widehat{\boldsymbol{\theta}}^{(m)}_j) = \underset{(\theta_{0,j},\boldsymbol{\theta}_j)}{\operatorname{argmin}}[-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; Z^*_{j,obs}, \mathbf{W}^{*(m)}_{j,obs}) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function.

$Z_{j,mis}$ is predicted with $Z^{(m)}_{j,mis}$ by drawing randomly from the predictive distribution $Bernoulli(\frac{exp(\widehat{\theta}^{(m)}_{0,j}\mathbf{1}_{n-r_j}+\mathbf{W}^{(m)}_{j,mis}\widehat{\boldsymbol{\theta}}^{(m)}_j)}{1+exp(\widehat{\theta}^{(m)}_{0,j}\mathbf{1}_{n-r_j}+\mathbf{W}^{(m)}_{j,mis}\widehat{\boldsymbol{\theta}}^{(m)}_j)})$. Let $Z^{(m)}_j = (Z^{(m)}_{j,mis}, Z_{j,obs})$.

(iii) If Z_j follows a Poisson distribution, the model is

$$log(\mathbf{E}[Z^*_{j,obs}|\mathbf{W}^{*(m)}_{j,obs}]) = \theta_{0,j}\mathbf{1}_{r^*_j} + \mathbf{W}^{*(m)}_{j,obs}\boldsymbol{\theta}_j, \tag{A.3}$$

A regularized regression method is used to fit model (A.3). The parameter estimates can be obtained as follows:

$$(\widehat{\theta}^{(m)}_{0,j}, \widehat{\boldsymbol{\theta}}^{(m)}_j) = \underset{(\theta_{0,j},\boldsymbol{\theta}_j)}{\operatorname{argmin}}[-\ell(\theta_{0,j}, \boldsymbol{\theta}_j; Z^*_{j,obs}, \mathbf{W}^{*(m)}_{j,obs}) + P_\lambda(\boldsymbol{\theta}_j)]$$

Where $P_\lambda(\boldsymbol{\theta}_j)$ is a regularization function.

$Z_{j,mis}$ is predicted with $Z^{(m)}_{j,mis}$ by drawing randomly from the predictive distribution $Poisson(exp(\widehat{\theta}^{(m)}_{0,j}\mathbf{1}_{n-r_j} + \mathbf{W}^{(m)}_{j,mis}\widehat{\boldsymbol{\theta}}^{(m)}_j))$. Let $Z^{(m)}_j = (Z^{(m)}_{j,mis}, Z_{j,obs})$.

We denote the updated data set after the m-th interation by $Z^{(m)}$ and repeat the procedures iteratively. After the algorithm converges, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.

A.2 Details of MICE-IURR for three types of data

We start the iterative procedure with some initial values. For example, all the elements in $Z_{mis,j}$ are filled in with the average of the observed values of Z_j $(j = 1, 2, ..., l)$. Define the corresponding initial completed dataset as $Z^{(0)}$.

In the m-th iteration:

(i) If Z_j follows a Gaussian distribution, we use a regularized regression method to fit a multiple linear regression model regarding $Z_{j,obs}$ as the outcome variable and $\mathbf{W}_{j,obs}^{(m)}$ as the predictor variable, and identify the active set, $\widehat{\mathcal{S}}_j^{(m)}$. Let $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ denote the subset of $\mathbf{W}_j^{(m)}$ that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$Z_{j,obs} = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}\boldsymbol{\theta}_j + \boldsymbol{\epsilon}_j, \tag{A.4}$$

where $\boldsymbol{\epsilon}_j \sim N(0, \sigma_j^2\mathbf{I}_{r_j})$ and $\mathbf{1}_{r_j}$ is a vector of length r_j with all entries one. Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j, \sigma_j^2)$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_{0,j}, \boldsymbol{\theta}_j, \sigma_j^2)' \sim N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$$

Where $\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}$ is the MLE of parameters in model (A.4) and $\widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)}$ is the variance-covariance matrix of the estimated parameters.
Generate a prediction for $Z_{j,mis}$: randomly draw $(\widehat{\theta}_{0,j}^{(m)}, \widehat{\boldsymbol{\theta}}_j^{(m)}, \widehat{\sigma}_j^{2(m)})$ from $N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$, and predict $Z_{j,mis}$ with $Z_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $N(\widehat{\theta}_{0,j}^{(m)}\mathbf{1}_{n-r_j} + \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}\widehat{\boldsymbol{\theta}}_j^{(m)}, \widehat{\sigma}_j^{2(m)}\mathbf{I}_{n-r_j})$. Let $Z_j^{(m)} = (Z_{j,mis}^{(m)}, Z_{j,obs})$.

(ii) If Z_j follows a Bernoulli distribution, we use a regularized regression method to fit a multiple linear regression model regarding $Z_{j,obs}$ as the outcome variable and $\mathbf{W}_{j,obs}^{(m)}$ as the predictor variable, and identify the active set, $\widehat{\mathcal{S}}_j^{(m)}$. Let $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ denote the subset of $\mathbf{W}_j^{(m)}$ that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$logit(\Pr(Z_{j,obs} = 1|\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs})) = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}\boldsymbol{\theta}_j, \tag{A.5}$$

Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j)$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_{0,j}, \boldsymbol{\theta}_j)' \sim N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$$

Where $\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}$ is the MLE of parameters in model (A.5) and $\widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)}$ is the variance-covariance matrix of the estimated parameters.

Generate a prediction for $Z_{j,mis}$: randomly draw $(\widehat{\theta}_{0,j}^{(m)}, \widehat{\boldsymbol{\theta}}_j^{(m)})$ from $N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$, and predict $Z_{j,mis}$ with $Z_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $Bernoulli(\frac{exp(\widehat{\theta}_{0,j}^{(m)}\mathbf{1}_{n-r_j}+\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}\widehat{\boldsymbol{\theta}}_j^{(m)})}{1+exp(\widehat{\theta}_{0,j}^{(m)}\mathbf{1}_{n-r_j}+\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}\widehat{\boldsymbol{\theta}}_j^{(m)})})$. Let $Z_j^{(m)} = (Z_{j,mis}^{(m)}, Z_{j,obs})$.

(iii) If Z_j follows a Poisson distribution, we use a regularized regression method to fit a multiple linear regression model regarding $Z_{j,obs}$ as the outcome variable and $\mathbf{W}_{j,obs}^{(m)}$ as the predictor variable, and identify the active set, $\widehat{\mathcal{S}}_j^{(m)}$. Let $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ denote the subset of $\mathbf{W}_j^{(m)}$ that only contains the active set. Correspondingly, denote two components of $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)}}$ by $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}$ and $\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}$. Then the model is

$$log(\mathbf{E}[Z_{j,obs}|\mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}]) = \theta_{0,j}\mathbf{1}_{r_j} + \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},obs}\boldsymbol{\theta}_j, \tag{A.6}$$

Approximate the distribution of $(\theta_{0,j}, \boldsymbol{\theta}_j)$ by using a standard inference procedure such as maximum likelihood.

$$(\theta_{0,j}, \boldsymbol{\theta}_j)' \sim N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$$

Where $\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}$ is the MLE of parameters in model (A.6) and $\widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)}$ is the variance-covariance matrix of the estimated parameters.

Generate a prediction for $Z_{j,mis}$: randomly draw $(\widehat{\theta}_{0,j}^{(m)}, \widehat{\boldsymbol{\theta}}_j^{(m)})$ from $N(\widehat{\boldsymbol{\theta}}_{MLE}^{(m)}, \widehat{\boldsymbol{\Sigma}}_{MLE}^{(m)})$, and predict $Z_{j,mis}$ with $Z_{j,mis}^{(m)}$ by drawing randomly from the predictive distribution $Poisson(exp(\widehat{\theta}_{0,j}^{(m)}\mathbf{1}_{n-r_j} + \mathbf{W}_{\widehat{\mathcal{S}}_j^{(m)},mis}\widehat{\boldsymbol{\theta}}_j^{(m)}))$. Let $Z_j^{(m)} = (Z_{j,mis}^{(m)}, Z_{j,obs})$.

We denote the updated data set after the m-th interation by $Z^{(m)}$ and repeat the procedures iteratively. After the algorithm converges, the last M imputed data sets after appropriate thinning are chosen for subsequent standard complete-data analysis.

Table A.1: Simulation results for estimating $\beta_1 = \beta_2 = \beta_3 = 1$ in the presence of missing data based on 200 monte carlo data sets, where $n = 100$ and $\rho = 0.1$. Bias, mean bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; KNN-V, KNN by nearest variables; KNN-S, KNN by nearest subjects; MICE-DURR, MICE through direct use of regularized regressions; MICE-IURR, MICE through indirect use of regularized regressions; EN, elastic net; Alasso, adaptive lasso.

		$\hat{\beta}_1$					$\hat{\beta}_2$					$\hat{\beta}_3$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
$q = 4$	GS	0.010	0.519	0.484	0.233	0.960	-0.004	0.518	0.514	0.263	0.930	-0.003	0.522	0.507	0.255	0.960
	CC	-0.273	0.560	0.570	0.398	0.910	-0.277	0.561	0.556	0.385	0.915	-0.350	0.566	0.570	0.446	0.865
	MI-true	0.023	0.718	0.676	0.455	0.940	-0.013	0.719	0.721	0.517	0.920	-0.109	0.728	0.692	0.489	0.950
							$p = 200$									
	MICE-RF	-0.618	0.697	0.364	0.514	0.970	-0.469	0.699	0.331	0.329	1.000	-0.615	0.701	0.341	0.494	0.940
	KNN-V	-0.800	0.696	0.584	0.979	0.845	-0.369	0.700	0.552	0.440	0.975	-0.696	0.712	0.547	0.782	0.900
	KNN-S	-0.273	0.560	0.570	0.398	0.910	-0.277	0.561	0.556	0.385	0.915	-0.350	0.566	0.570	0.446	0.865
	MICE-DURR(Lasso)	-0.781	0.572	0.404	0.773	0.775	-0.543	0.567	0.431	0.479	0.895	-0.759	0.583	0.431	0.760	0.775
	MICE-DURR(EN)	-0.784	0.575	0.407	0.779	0.790	-0.534	0.568	0.428	0.467	0.895	-0.759	0.585	0.432	0.761	0.755
	MICE-DURR(Alasso)	-0.774	0.589	0.374	0.739	0.800	-0.557	0.581	0.384	0.458	0.925	-0.768	0.600	0.392	0.742	0.820
	MICE-IURR(Lasso)	-0.031	0.684	0.711	0.503	0.935	-0.017	0.681	0.697	0.484	0.920	-0.173	0.692	0.676	0.485	0.930
	MICE-IURR(EN)	-0.047	0.673	0.731	0.534	0.920	-0.009	0.670	0.703	0.491	0.910	-0.144	0.676	0.703	0.513	0.900
	MICE-IURR(Alasso)	-0.105	0.759	0.813	0.668	0.905	-0.181	0.751	0.794	0.660	0.910	-0.283	0.759	0.751	0.641	0.925
							$p = 1000$									
	MICE-DURR(Lasso)	-0.769	0.579	0.412	0.761	0.810	-0.568	0.576	0.474	0.547	0.885	-0.720	0.580	0.405	0.682	0.855
	MICE-DURR(EN)	-0.766	0.578	0.414	0.758	0.835	-0.558	0.581	0.465	0.526	0.900	-0.719	0.584	0.413	0.687	0.840
	MICE-DURR(Alasso)	-0.748	0.582	0.364	0.691	0.855	-0.576	0.583	0.441	0.525	0.905	-0.751	0.588	0.396	0.720	0.815
	MICE-IURR(Lasso)	0.021	0.683	0.780	0.605	0.885	-0.037	0.677	0.778	0.603	0.910	-0.168	0.689	0.714	0.536	0.910
	MICE-IURR(EN)	0.050	0.671	0.776	0.601	0.890	-0.049	0.671	0.779	0.606	0.900	-0.158	0.680	0.724	0.546	0.915
	MICE-IURR(Alasso)	-0.159	0.791	0.901	0.834	0.905	-0.245	0.777	0.849	0.777	0.905	-0.366	0.786	0.767	0.720	0.925
$q = 20$	GS	0.012	0.540	0.492	0.241	0.975	0.010	0.535	0.548	0.299	0.930	-0.008	0.542	0.508	0.256	0.965
	CC	-0.366	0.529	0.554	0.439	0.865	-0.359	0.526	0.517	0.395	0.905	-0.443	0.535	0.495	0.440	0.865
	MI-true	-0.097	0.715	0.627	0.400	0.950	-0.089	0.713	0.660	0.441	0.930	-0.133	0.728	0.683	0.482	0.950
							$p = 200$									
	MICE-RF	-0.468	0.730	0.429	0.402	0.975	-0.384	0.710	0.416	0.320	0.980	-0.447	0.725	0.412	0.369	0.980
	KNN-V	-0.412	0.684	0.527	0.446	0.945	-0.266	0.678	0.531	0.351	0.960	-0.389	0.691	0.544	0.445	0.945
	KNN-S	-0.366	0.529	0.554	0.439	0.865	-0.359	0.526	0.517	0.395	0.905	-0.443	0.535	0.495	0.440	0.865
	MICE-DURR(Lasso)	-0.494	0.675	0.464	0.458	0.960	-0.374	0.666	0.442	0.334	0.970	-0.443	0.678	0.442	0.390	0.970
	MICE-DURR(EN)	-0.499	0.677	0.480	0.478	0.965	-0.379	0.660	0.430	0.328	0.965	-0.435	0.679	0.448	0.389	0.975
	MICE-DURR(Alasso)	-0.488	0.685	0.475	0.462	0.945	-0.378	0.675	0.412	0.312	0.980	-0.457	0.693	0.423	0.387	0.985
	MICE-IURR(Lasso)	-0.058	0.695	0.714	0.510	0.915	0.058	0.671	0.732	0.536	0.875	-0.048	0.698	0.756	0.570	0.905
	MICE-IURR(EN)	-0.049	0.691	0.705	0.496	0.930	0.023	0.679	0.715	0.509	0.880	-0.041	0.697	0.743	0.551	0.910
	MICE-IURR(Alasso)	-0.214	0.714	0.735	0.583	0.905	-0.197	0.700	0.774	0.635	0.890	-0.309	0.719	0.760	0.671	0.880
							$p = 1000$									
	MICE-DURR(Lasso)	-0.435	0.677	0.470	0.409	0.970	-0.401	0.673	0.461	0.372	0.964	-0.471	0.679	0.437	0.412	0.976
	MICE-DURR(EN)	-0.443	0.673	0.476	0.422	0.964	-0.403	0.669	0.476	0.387	0.952	-0.451	0.678	0.463	0.416	0.982
	MICE-DURR(Alasso)	-0.434	0.678	0.475	0.412	0.976	-0.401	0.681	0.451	0.363	0.982	-0.474	0.683	0.433	0.411	0.970
	MICE-IURR(Lasso)	0.095	0.686	0.781	0.616	0.909	-0.019	0.687	0.858	0.732	0.885	-0.073	0.711	0.737	0.545	0.933
	MICE-IURR(EN)	0.082	0.689	0.776	0.606	0.897	-0.009	0.687	0.870	0.752	0.867	-0.069	0.705	0.749	0.562	0.915
	MICE-IURR(Alasso)	-0.212	0.739	0.758	0.615	0.927	-0.352	0.721	0.767	0.708	0.897	-0.394	0.732	0.745	0.707	0.885

Table A.2: Simulation results for estimating $\beta_1 = \beta_2 = \beta_3 = 1$ in the presence of missing data based on 200 monte carlo data sets, where $n = 100$ and $\rho = 0.5$. Bias, mean bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; KNN-V, KNN by nearest variables; KNN-S, KNN by nearest subjects; MICE-DURR, MICE through direct use of regularized regressions; MICE-IURR, MICE through indirect use of regularized regressions; EN, elastic net; Alasso, adaptive lasso.

		$\hat{\beta}_1$					$\hat{\beta}_2$					$\hat{\beta}_3$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
$q = 4$	GS	0.007	0.515	0.478	0.228	0.970	-0.014	0.515	0.533	0.283	0.940	-0.037	0.519	0.496	0.246	0.970
	CC	-0.274	0.569	0.560	0.387	0.890	-0.272	0.570	0.588	0.418	0.910	-0.265	0.573	0.529	0.349	0.935
	MI-true	-0.068	0.687	0.687	0.474	0.925	0.010	0.698	0.717	0.512	0.915	-0.062	0.693	0.685	0.470	0.925
							$p = 200$									
	MICE-RF	-0.628	0.747	0.352	0.518	0.960	-0.355	0.711	0.371	0.263	0.980	-0.510	0.725	0.349	0.382	0.975
	KNN-V	-0.930	0.708	0.545	1.160	0.800	-0.242	0.711	0.535	0.343	0.980	-0.655	0.720	0.557	0.738	0.925
	KNN-S	-0.274	0.569	0.560	0.387	0.890	-0.272	0.570	0.588	0.418	0.910	-0.265	0.573	0.529	0.349	0.935
	MICE-DURR(Lasso)	-0.853	0.548	0.407	0.892	0.720	-0.475	0.544	0.384	0.373	0.925	-0.688	0.546	0.418	0.647	0.860
	MICE-DURR(EN)	-0.852	0.548	0.412	0.894	0.695	-0.474	0.541	0.386	0.372	0.935	-0.688	0.550	0.411	0.641	0.845
	MICE-DURR(Alasso)	-0.844	0.566	0.349	0.834	0.785	-0.486	0.563	0.351	0.359	0.945	-0.701	0.570	0.374	0.631	0.870
	MICE-IURR(Lasso)	-0.094	0.668	0.690	0.483	0.930	-0.080	0.669	0.729	0.536	0.885	-0.068	0.661	0.703	0.496	0.925
	MICE-IURR(EN)	-0.092	0.653	0.701	0.498	0.935	-0.074	0.667	0.724	0.527	0.910	-0.069	0.647	0.705	0.499	0.915
	MICE-IURR(Alasso)	-0.091	0.720	0.744	0.559	0.940	-0.076	0.710	0.790	0.626	0.880	-0.123	0.719	0.761	0.592	0.920
							$p = 1000$									
	MICE-DURR(Lasso)	-0.809	0.560	0.392	0.807	0.715	-0.534	0.558	0.445	0.482	0.905	-0.713	0.568	0.437	0.698	0.825
	MICE-DURR(EN)	-0.806	0.560	0.398	0.808	0.710	-0.530	0.560	0.458	0.490	0.920	-0.711	0.568	0.433	0.693	0.815
	MICE-DURR(Alasso)	-0.803	0.567	0.354	0.770	0.755	-0.544	0.565	0.409	0.463	0.905	-0.730	0.577	0.379	0.676	0.840
	MICE-IURR(Lasso)	0.058	0.677	0.836	0.698	0.860	-0.040	0.676	0.811	0.656	0.890	-0.197	0.677	0.752	0.601	0.905
	MICE-IURR(EN)	0.073	0.664	0.808	0.655	0.850	-0.041	0.667	0.800	0.638	0.890	-0.192	0.670	0.739	0.581	0.885
	MICE-IURR(Alasso)	-0.036	0.793	0.912	0.829	0.865	-0.209	0.783	0.852	0.765	0.885	-0.335	0.779	0.792	0.737	0.910
$q = 20$	GS	0.014	0.521	0.500	0.249	0.950	-0.008	0.523	0.550	0.301	0.920	-0.023	0.527	0.506	0.255	0.970
	CC	-0.352	0.525	0.536	0.410	0.855	-0.357	0.529	0.541	0.418	0.875	-0.387	0.533	0.520	0.419	0.890
	MI-true	-0.077	0.689	0.593	0.356	0.965	-0.047	0.688	0.675	0.456	0.920	-0.092	0.692	0.628	0.401	0.965
							$p = 200$									
	MICE-RF	-0.390	0.717	0.427	0.333	0.990	-0.286	0.708	0.466	0.298	0.985	-0.412	0.721	0.382	0.315	0.995
	KNN-V	-0.450	0.685	0.530	0.482	0.950	-0.214	0.697	0.561	0.359	0.970	-0.441	0.703	0.469	0.413	0.980
	KNN-S	-0.352	0.525	0.536	0.410	0.855	-0.357	0.529	0.541	0.418	0.875	-0.387	0.533	0.520	0.419	0.890
	MICE-DURR(Lasso)	-0.491	0.631	0.451	0.444	0.955	-0.329	0.633	0.486	0.343	0.970	-0.532	0.644	0.421	0.459	0.975
	MICE-DURR(EN)	-0.489	0.631	0.455	0.445	0.920	-0.340	0.627	0.486	0.351	0.970	-0.518	0.641	0.401	0.428	0.970
	MICE-DURR(Alasso)	-0.481	0.655	0.425	0.412	0.940	-0.348	0.653	0.452	0.325	0.975	-0.529	0.668	0.377	0.422	0.960
	MICE-IURR(Lasso)	-0.022	0.678	0.729	0.530	0.905	-0.024	0.684	0.752	0.563	0.870	-0.057	0.683	0.689	0.475	0.905
	MICE-IURR(EN)	-0.023	0.666	0.695	0.481	0.920	-0.022	0.670	0.730	0.531	0.855	-0.036	0.675	0.674	0.453	0.925
	MICE-IURR(Alasso)	-0.144	0.731	0.721	0.539	0.910	-0.103	0.719	0.764	0.591	0.895	-0.181	0.720	0.718	0.546	0.955
							$p = 1000$									
	MICE-DURR(Lasso)	-0.436	0.640	0.415	0.361	0.978	-0.399	0.628	0.490	0.396	0.956	-0.458	0.640	0.393	0.363	0.989
	MICE-DURR(EN)	-0.459	0.631	0.437	0.399	0.978	-0.375	0.626	0.492	0.380	0.967	-0.460	0.636	0.407	0.375	1.000
	MICE-DURR(Alasso)	-0.450	0.651	0.419	0.376	0.978	-0.387	0.642	0.482	0.380	0.978	-0.487	0.655	0.384	0.382	0.989
	MICE-IURR(Lasso)	0.012	0.682	0.750	0.556	0.900	0.004	0.660	0.712	0.502	0.956	-0.079	0.687	0.630	0.398	0.933
	MICE-IURR(EN)	0.020	0.672	0.738	0.539	0.878	0.018	0.662	0.725	0.521	0.911	-0.086	0.653	0.660	0.438	0.933
	MICE-IURR(Alasso)	-0.204	0.766	0.767	0.624	0.878	-0.135	0.745	0.778	0.617	0.889	-0.318	0.737	0.666	0.540	0.911

Table A.3: Simulation results for estimating $\beta_1 = \beta_2 = \beta_3 = 1$ in the presence of missing data based on 200 monte carlo data sets, where $n = 100$ and $\rho = 0.9$. Bias, mean bias; SE, mean standard error; SD, Monte Carlo standard deviation; MSE, mean square error; CR, coverage rate of 95% confidence interval; GS, gold standard; CC, complete-case; KNN-V, KNN by nearest variables; KNN-S, KNN by nearest subjects; MICE-DURR, MICE through direct use of regularized regressions; MICE-IURR, MICE through indirect use of regularized regressions; EN, elastic net; Alasso, adaptive lasso.

		$\hat{\beta}_1$					$\hat{\beta}_2$					$\hat{\beta}_3$				
		Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR	Bias	SE	SD	MSE	CR
$q = 4$	GS	0.029	0.515	0.476	0.226	0.960	0.003	0.512	0.521	0.270	0.945	0.008	0.517	0.489	0.238	0.970
	CC	-0.219	0.583	0.609	0.417	0.905	-0.245	0.585	0.560	0.371	0.925	-0.242	0.582	0.582	0.395	0.945
	MI-true	-0.012	0.711	0.699	0.487	0.935	-0.027	0.710	0.715	0.509	0.915	-0.007	0.701	0.685	0.467	0.940
							$p = 200$									
	MICE-RF	-0.483	0.721	0.358	0.361	0.980	-0.277	0.704	0.376	0.218	1.000	-0.372	0.722	0.380	0.283	0.975
	KNN-V	-1.034	0.747	0.584	1.408	0.765	-0.113	0.741	0.596	0.366	0.990	-0.676	0.758	0.630	0.852	0.895
	KNN-S	-0.219	0.583	0.609	0.417	0.905	-0.245	0.585	0.560	0.371	0.925	-0.242	0.582	0.582	0.395	0.945
	MICE-DURR(Lasso)	-0.857	0.547	0.435	0.922	0.670	-0.549	0.532	0.417	0.475	0.905	-0.643	0.545	0.441	0.606	0.835
	MICE-DURR(EN)	-0.862	0.548	0.443	0.939	0.670	-0.534	0.537	0.420	0.461	0.880	-0.649	0.550	0.438	0.612	0.815
	MICE-DURR(Alasso)	-0.841	0.571	0.409	0.874	0.745	-0.525	0.564	0.377	0.417	0.935	-0.677	0.576	0.396	0.615	0.855
	MICE-IURR(Lasso)	-0.049	0.676	0.701	0.491	0.915	-0.047	0.671	0.694	0.481	0.920	-0.027	0.669	0.702	0.492	0.920
	MICE-IURR(EN)	-0.028	0.671	0.716	0.511	0.915	-0.059	0.676	0.707	0.501	0.910	-0.052	0.677	0.706	0.498	0.935
	MICE-IURR(Alasso)	-0.005	0.684	0.710	0.502	0.920	-0.024	0.685	0.727	0.526	0.915	-0.031	0.679	0.710	0.503	0.935
							$p = 1000$									
	MICE-DURR(Lasso)	-0.787	0.556	0.399	0.778	0.740	-0.494	0.538	0.405	0.408	0.920	-0.688	0.559	0.375	0.613	0.845
	MICE-DURR(EN)	-0.793	0.554	0.414	0.799	0.715	-0.490	0.534	0.410	0.407	0.900	-0.681	0.562	0.373	0.603	0.840
	MICE-DURR(Alasso)	-0.801	0.572	0.363	0.772	0.775	-0.496	0.557	0.369	0.382	0.930	-0.705	0.577	0.339	0.611	0.880
	MICE-IURR(Lasso)	-0.014	0.682	0.744	0.550	0.930	-0.058	0.672	0.763	0.582	0.915	-0.144	0.680	0.709	0.521	0.920
	MICE-IURR(EN)	0.009	0.679	0.729	0.529	0.935	-0.062	0.667	0.744	0.555	0.915	-0.155	0.672	0.702	0.515	0.930
	MICE-IURR(Alasso)	-0.029	0.729	0.739	0.545	0.935	-0.024	0.716	0.818	0.666	0.895	-0.130	0.715	0.760	0.592	0.910
$q = 20$	GS	0.018	0.514	0.478	0.228	0.945	-0.010	0.512	0.519	0.268	0.945	-0.003	0.515	0.488	0.237	0.975
	CC	-0.260	0.570	0.562	0.382	0.930	-0.295	0.566	0.533	0.370	0.930	-0.304	0.568	0.537	0.379	0.925
	MI-true	-0.056	0.681	0.561	0.316	0.975	-0.075	0.671	0.590	0.352	0.980	-0.009	0.667	0.566	0.319	0.960
							$p = 200$									
	MICE-RF	-0.334	0.747	0.361	0.241	0.995	-0.153	0.707	0.362	0.154	1.000	-0.269	0.756	0.398	0.230	1.000
	KNN-V	-0.798	0.749	0.555	0.943	0.895	-0.108	0.739	0.555	0.318	1.000	-0.506	0.757	0.569	0.578	0.960
	KNN-S	-0.260	0.570	0.562	0.382	0.930	-0.295	0.566	0.533	0.370	0.930	-0.304	0.568	0.537	0.379	0.925
	MICE-DURR(Lasso)	-0.690	0.576	0.412	0.645	0.880	-0.392	0.557	0.404	0.316	0.945	-0.575	0.574	0.406	0.495	0.930
	MICE-DURR(EN)	-0.679	0.579	0.420	0.636	0.880	-0.379	0.556	0.404	0.306	0.955	-0.591	0.572	0.414	0.520	0.920
	MICE-DURR(Alasso)	-0.668	0.616	0.382	0.592	0.915	-0.393	0.593	0.391	0.306	0.970	-0.611	0.611	0.381	0.519	0.945
	MICE-IURR(Lasso)	-0.051	0.665	0.646	0.418	0.935	-0.031	0.666	0.678	0.458	0.910	0.007	0.660	0.658	0.431	0.950
	MICE-IURR(EN)	-0.031	0.665	0.645	0.415	0.935	-0.057	0.662	0.660	0.436	0.930	0.006	0.659	0.655	0.427	0.945
	MICE-IURR(Alasso)	-0.035	0.668	0.691	0.476	0.915	-0.063	0.667	0.660	0.437	0.935	-0.003	0.667	0.684	0.466	0.935
							$p = 1000$									
	MICE-DURR(Lasso)	-0.738	0.582	0.417	0.717	0.850	-0.408	0.569	0.417	0.340	0.955	-0.531	0.586	0.444	0.478	0.905
	MICE-DURR(EN)	-0.729	0.587	0.416	0.704	0.860	-0.416	0.569	0.421	0.350	0.945	-0.526	0.587	0.443	0.472	0.925
	MICE-DURR(Alasso)	-0.723	0.614	0.384	0.670	0.920	-0.433	0.600	0.374	0.327	0.950	-0.540	0.612	0.401	0.451	0.945
	MICE-IURR(Lasso)	0.055	0.687	0.725	0.526	0.905	-0.057	0.676	0.770	0.593	0.890	-0.124	0.676	0.735	0.553	0.895
	MICE-IURR(EN)	0.052	0.675	0.746	0.556	0.900	-0.053	0.673	0.782	0.611	0.910	-0.113	0.669	0.751	0.573	0.915
	MICE-IURR(Alasso)	-0.003	0.742	0.769	0.589	0.930	-0.105	0.728	0.803	0.653	0.920	-0.082	0.732	0.792	0.631	0.915

Bibliography

Allison, P. D. (2001), *Missing data*, Vol. 136, Sage publications.

Audigier, V., Husson, F. and Josse, J. (2016), ‘A principal component method to impute missing values for mixed data’, *Advances in Data Analysis and Classification* **10**(1), 5–26.

Boscardin, W. J., Zhang, X. and Belin, T. R. (2008), ‘Modeling a mixture of ordinal and continuous repeated measures’, *Journal of Statistical Computation and Simulation* **78**(10), 873–886.

Brakerski, Z. (2012), Fully homomorphic encryption without modulus switching from classical gapsvp, *in* ‘Advances in Cryptology–CRYPTO 2012’, Springer, pp. 868–886.

Breiman, L. (2001), ‘Random forests’, *Machine learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

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

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**(3), 259–275.

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**.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E. et al. (2013), 'Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data', *Nature biotechnology* **31**(12), 1102–1111.

Du, W., Han, Y. S. and Chen, S. (2004), Privacy-preserving multivariate statistical analysis: Linear regression and classification., *in* 'SDM', Vol. 4, SIAM, pp. 222–233.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. and Erlich, Y. (2013), 'Identifying personal genomes by surname inference', *Science* **339**(6117), 321–324.

Haitovsky, Y. (1968), 'Missing data in regression analysis', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 67–82.

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

He, R. and Belin, T. (2014), 'Multiple imputation for high-dimensional mixed incomplete continuous and binary data', *Statistics in medicine* **33**(13), 2251–2262.

He, Y., Yucel, R. and Raghunathan, T. (2011), 'A functional multiple imputation approach to incomplete longitudinal data', *Statistics in Medicine* **30**(10), 1137–1156.

Höfler, M., Pfister, H., Lieb, R. and Wittchen, H.-U. (2005), 'The use of weights to account for non-response and drop-out', *Social psychiatry and psychiatric epidemiology* **40**(4), 291–299.

Holan, S. H., Toth, D., Ferreira, M. A. R. and Karr, A. F. (2010), 'Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality', *Journal of the American Statistical Association* **105**(490), 564–577.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. and Craig, D. W. (2008), 'Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays', PLoS Genet **4**(8), e1000167.

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

Hothorn, T., Hornik, K. and Zeileis, A. (2006), 'Unbiased recursive partitioning: A conditional inference framework', Journal of Computational and Graphical statistics **15**(3), 651–674.

Hsu, C., Taylor, J., Murray, S. and Commenges, D. (2004), 'Survival analysis using auxiliary variables via nonparametric multiple imputation'.

Jaakkola, T. S. and Haussler, D. (1999), 'Probabilistic kernel regression models.', AISTATS .

Jagannathan, G. and Wright, R. N. (2008), 'Privacy-preserving imputation of missing data', Data & Knowledge Engineering **65**(1), 40–56.

Jiang, W., Li, P., Wang, S., Wu, Y., Xue, M., Ohno-Machado, L. and Jiang, X. (2013), 'WebGLORE: a Web service for Grid LOgistic REgression', Bioinformatics **29**(24), 3238–3240.

Jiang, X., Sarwate, A. D. and Ohno-Machado, L. (2013), 'Privacy technology to support data sharing for comparative effectiveness research: a systematic review', Medical care **51**(8 0 3), S58.

Josse, J., Pagès, J. and Husson, F. (2011), 'Multiple imputation in principal component analysis', Advances in data analysis and classification **5**(3), 231–246.

Karr, A. F., Lin, X., Sanil, A. P. and Reiter, J. P. (2009), ‘Privacy-preserving analysis of vertically partitioned data using secure matrix products’, Journal of Official Statistics **25**(1), 125.

Li, K.-H. (1988), ‘Imputation using markov chains’, Journal of Statistical Computation and Simulation **30**(1), 57–79.

Li, Y., Jiang, X., Wang, S., Xiong, H. and Ohno-Machado, L. (2015), ‘Vertical grid logistic regression (vertigo)’, Journal of the American Medical Informatics Association p. ocv146.

Li, Y., Jiang, X., Wang, S., Xiong, H. and Ohno-Machado, L. (2016), ‘VERTIcal Grid lOgistic regression (VERTIGO)’, Journal of the American Medical Informatics Association **23**(3), 570–579.

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

Little, R. and An, H. (2004), ‘Robust likelihood-based analysis of multivariate data with missing values’, Statistica Sinica **14**, 949–968.

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

Little, R. and Rubin, D. (2002), Statistical Analysis with Missing Data, John Wiley: New York.

Liu, J., Gelman, A., Hill, J., Su, Y.-S. and Kropko, J. (2013), ‘On the stationary distribution of iterative imputations’, Biometrika p. ast044.

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

Marimont, R. and Shapiro, M. (1979), 'Nearest neighbour searches and the curse of dimensionality', IMA Journal of Applied Mathematics **24**(1), 59–70.

Maro, J. C., Platt, R., Holmes, J. H., Strom, B. L., Hennessy, S., Lazarus, R. and Brown, J. S. (2009), 'Design of a national distributed health data network', Annals of internal medicine **151**(5), 341–344.

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

Millsap, R. E. and Maydeu-Olivares, A. (2009), The SAGE handbook of quantitative methods in psychology, Sage Publications.

Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.-P., Malin, B., Wang, X. et al. (2014), 'Privacy and security in the genomic era'.

Newton, K. M., Peissig, P. L., Kho, A. N., Bielinski, S. J., Berg, R. L., Choudhary, V., Basford, M., Chute, C. G., Kullo, I. J., Li, R. et al. (2013), 'Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network', Journal of the American Medical Informatics Association **20**(e1), e147–e154.

Nielsen, S. F. (2003), 'Proper and improper multiple imputation', International Statistical Review **71**(3), 593–607.

Peng, Z., Yan, M. and Yin, W. (2013), Parallel and distributed sparse optimization, in '2013 Asilomar Conference on Signals, Systems and Computers', IEEE, pp. 659–646.

Penny, K. I. and Atkinson, I. (2012), 'Approaches for dealing with missing data in health care studies', Journal of clinical nursing **21**(19pt20), 2722–2729.

Powell, M. (1964), 'An efficient method for finding the minimum of a function of several variables without calculating derivatives', The computer journal .

Qi, L., Wang, Y.-F. and He, Y. (2010), 'A comparison of multiple imputation and fully augmented weighted estimators for cox regression with missing covariates', Statistics in medicine **29**(25), 2592–2604.

Raghunathan, T. E. and Siscovick, D. S. (1996), 'A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives', Applied Statistics pp. 335–352.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', Survey methodology **27**(1), 85–96.

Raghupathi, W. and Raghupathi, V. (2014), 'Big data analytics in healthcare: promise and potential', Health Information Science and Systems **2**(1), 211–10.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', Journal of the American statistical Association **89**(427), 846–866.

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

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

Sanil, A. P., Karr, A. F., Lin, X. and Reiter, J. P. (2004), Privacy preserving regression modelling via distributed computation, in 'Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 677–682.

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

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.

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

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012), ‘Combining multiple imputation and inverse-probability weighting’, Biometrics **68**(1), 129–137.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and Hemingway, H. (2014), ‘Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study’, American Journal of Epidemiology **179**(6), 764–774.

Silva-Ramírez, E.-L., Pino-Mejías, R. and López-Coello, M. (2015), ‘Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns’, Applied Soft Computing **29**, 65–74.

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M. and Cubiles-de-la Vega, M.-D. (2011), ‘Missing value imputation on missing completely at random data using multilayer perceptrons’, Neural Networks **24**(1), 121–129.

Slavkovic, A. B., Nardi, Y. and Tibbits, M. M. (2007), ” secure” logistic regression of horizontally and vertically partitioned distributed databases, in ‘Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)’, IEEE, pp. 723–728.

Song, J. and Belin, T. R. (2004), ‘Imputation for incomplete high-dimensional multivariate normal data using a common factor model’, Statistics in medicine **23**(18), 2827–2843.

Städler, N. and Bühlmann, P. (2012), ‘Missing values: sparse inverse covariance estimation and an extension to sparse regression’, Statistics and Computing **22**(1), 219–235.

Städler, N., Stekhoven, D. J. and Bühlmann, P. (2014), ‘Pattern alternating maximization algorithm for missing data in high-dimensional problems.’, Journal of Machine Learning Research **15**(1), 1903–1928.

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

Stone, C. J. (1980), 'Optimal rates of convergence for nonparametric estimators', The annals of Statistics **8**(6), 1348–1360.

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

Su, Y.-S., Gelman, A., Hill, J., Yajima, M. et al. (2011), 'Multiple imputation with diagnostics (mi) in r: Opening windows into the black box', Journal of Statistical Software **45**(2), 1–31.

Tanner, M. A. and Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', Journal of the American statistical Association **82**(398), 528–540.

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

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001), 'Missing value estimation methods for dna microarrays', Bioinformatics **17**(6), 520–525.

Tutz, G. and Ramzan, S. (2015), 'Improved methods for the imputation of missing data by nearest neighbor methods', Computational Statistics & Data Analysis **90**, 84–99.

Vach, W. (2012), Logistic regression with missing values in the covariates, Vol. 86, Springer Science & Business Media.

Vaidya, J. and Clifton, C. (2002), 'Privacy preserving association rule mining in vertically partitioned data', ... on Knowledge discovery and data mining .

Vaidya, J. and Clifton, C. (2003), Privacy-preserving k-means clustering over vertically partitioned data, in 'Proceedings of the ninth ACM SIGKDD international ...'.

Vaidya, J. and Clifton, C. (2004), Privacy preserving naive bayes classifier for vertically partitioned data, in 'Proceedings of the 2004 SIAM International ...', Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 522–526.

Valdiviezo, H. C. and Van Aelst, S. (2015), 'Tree-based prediction on incomplete data using imputation or surrogate decisions', Information Sciences **311**, 163–181.

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

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

Vansteelandt, S., Carpenter, J. and Kenward, M. G. (2010), 'Analysis of incomplete data using inverse probability weighting and doubly robust estimators', Methodology .

Wang, C. Y., Wang, S., Zhao, L. P. and Ou, S. T. (1997), 'Weighted semiparametric estimation in regression analysis with missing covariate data', Journal of the American ... **92**(438), 512–525.

Wang, R., Li, Y. F., Wang, X., Tang, H. and Zhou, X. (2009), Learning your identity and disease from research papers: information leaks in genome wide association study, in 'Proceedings of the 16th ACM conference on Computer and communications security', ACM, pp. 534–544.

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

Willenborg, L. and de Waal, T. (2012), Elements of Statistical Disclosure Control, Springer Science & Business Media.

Wu, Y., Jiang, X., Kim, J. and Ohno-Machado, L. (2012), 'Grid binary logistic regression (glore): building shared models without sharing data', Journal of the American Medical Informatics Association **19**(5), 758–764.

Yin, X., Levy, D., Willinger, C., Adourian, A. and Larson, M. G. (2016), 'Multiple imputation and analysis for high-dimensional incomplete proteomics data', Statistics in medicine **35**(8), 1315–1326.

Yu, H., Vaidya, J. and Jiang, X. (2006), 'Privacy-Preserving SVM Classification on Vertically Partitioned Data.', PAKDD .

Zhang, G. and Little, R. (2008), 'Extensions of the penalized spline of propensity prediction method of imputation', Biometrics **65**(3), 911–918.

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

Zhu, J. and Raghunathan, T. E. (2014), 'Convergence properties of a sequential regression multiple imputation algorithm', Journal of the American Statistical Association (just-accepted), 00–00.

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

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320.