Projective Resampling Imputation Mean Estimation Method for Missing Covariates Problem

Zishu Zhan\textsuperscript{1,2}, Xiangjie Li\textsuperscript{3,*}, Jingxiao Zhang\textsuperscript{1,2,*}

June 17, 2021

\textsuperscript{1}Center for Applied Statistics, Renmin University of China,  
\textsuperscript{2}School of Statistics, Remin University of China,  
\textsuperscript{3}State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Abstract

Missing data is a common problem in clinical data collection, which causes difficulty in the statistical analysis of such data. To overcome problems caused by incomplete data, we propose a new imputation method called projective resampling imputation mean estimation (PRIME), which can also address “the curse of dimensionality” problem in imputation with less information loss. We use various sample sizes, missing-data rates, covariate correlations, and noise levels in simulation studies, and all results show that PRIME outperforms other methods such as iterative least-squares estimation (ILSE), maximum likelihood (ML), and complete-case analysis (CC). Moreover, we conduct a study of influential factors in cardiac surgery-associated acute kidney injury (CSA-AKI), which show that our method performs better than the other models. Finally, we prove that PRIME has a consistent property under some regular conditions.

Keywords: projective resampling, missing covariates problem, imputation method, linear regression

1 Introduction

In medical research, an investigator’s ultimate interest may be in inferring prognostic markers, given the patients’ genetic, cytokine, and/or environmental backgrounds \cite{1, 2}. However, in practical applications, data are often missing. The most common approach to address missing-data problems is complete-case analysis (CC), which is simple but inefficient. CC can also lead to biased estimates when the data are not missing completely at random. The maximum likelihood method (ML \cite{3, 4}) and the inverse probability weighting method (IPW \cite{5, 6}) are also widely used approaches to address missing data. However, likelihood-based methods are sensitive to model assumptions, and re-weighting methods do not always make full use of the
available data. Alternatively, imputation [7, 8] is a more flexible approach to couple with missing data.

However, a preliminary analysis of cardiac surgery-associated acute kidney injury (CSA-AKI) data used in Chen et al. [2] indicates that the missing-data patterns vary across individuals. Accordingly, new and more capable quantitative methods are needed for this individual-specific missing data. Furthermore, it is common for only a small fraction of records to have complete information across all sources. Existing methods do not work well when the percentage of unavailable data is high. To estimate the coefficients (rather than predicting them), Lin et al. [9] proposed the iterative least-squares estimation (ILSE) method to deal with individual-specific missing-data patterns using the classical regression framework, but it needs a complete set of observations to obtain the initial values, and its results might not converge when based on bad initial values. Furthermore, Lin et al. [9] may have difficulty accommodating data missing from both important and unimportant variables.

In this study, based on the idea of projection resampling/random projection, we propose Projection Resampling Imputation Mean Estimation (PRIME), a method that tackles the aforementioned drawbacks of existing methods. The key idea behind projection resampling/random projection was given in the Johnson-Lindenstrauss lemma [10], which preserves pairwise distances after projecting a set of points to a randomly chosen low-dimensional subspace. There are several previous studies on projection resampling/random projection for dimension reduction, including Schulman [11] for clustering, Donoho [12] for signal processing, Shi et al. [13] for classification, Maillard and Munos [14] for linear regression, and Le et al. [15] for kernel approximation. Specifically, the idea of PRIME is to project the covariates along randomly sampled directions to obtain samples of scalar-valued predictors and kernels (dimension reduction). Next, a simple geometric average is taken on the scalar-predictor-based kernel to impute the missing parts (using all-sided information). Our method has several advantages, including the following. First, PRIME can deal with a high degree of missing data, even data containing no complete observations, while most existing methods require at least a fraction of the subjects to have fully complete observations. Second, we can average the imputed estimates from multiple projection directions and fully utilize the available information to reach a more reliable and useful result. Third, to reduce the undesirable influence of unimportant variables, PRIME can be easily extended to sparse PRIME (denoted as SPRIME), which has a profound impact in practical applications.

The remainder of the paper is organized as follows. Section 2 introduces the basic setup of PRIME and SPRIME. Theoretical properties are discussed in Section 3. Sections 4 and 5 present the numerical results using simulated and real data examples, respectively. Section 6 presents some concluding remarks. In addition, our proposed method is implemented using R and the scripts to reproduce our results are available at https://github.com/eleozrr/PRIME. The proof of the theorem is available the Appendix.
2 Projective resampling Imputation mean estimation (PRIME)

2.1 Model and estimation by PRIME

In this paper, let \( Y = (Y_1, Y_2, \ldots, Y_n)^T \) be the response variable of interest and \( X = \{X_{ij} : i = 1, 2, \ldots, n, j = 1, 2, \ldots, p \} \) be the covariate matrix. We assume that \( n > p \). We consider the presence of missing covariates by \( r_{ij} \), which denotes the missing-data indicator for \( X_{ij} \), where \( r_{ij} = 1 \) if \( X_{ij} \) is missing and is 0 otherwise. For each unit \( i \), \( A_i = \{j : r_{ij} = 0, j = 1, 2, \ldots, p \} \) denotes the available covariates set, e.g., \( A_i = \{1, 2, \ldots, p \} \) for the complete case.

In this study, we focus on a linear regression model. Assume that the random sample \( \{(Y_i, X_i) : i = 1, 2, \ldots, n, j = 1, 2, \ldots, p \} \) is generated by:

\[
Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i = X_i^T \beta + \varepsilon_i, \tag{1}
\]

where \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \) is the coefficient vector for the covariates and the \( \varepsilon_i \)'s are independently identically distributed random errors. We assume \( E(\varepsilon_i | X_i) = 0 \) and \( E(\varepsilon_i^2 | X_i) = \sigma^2 \).

Covariates can be divided into two parts based on \( A_i \): \( X_{i,A_i} = (X_{ij} : j \in A_i)^T \) for observed covariates and \( X_{i,\bar{A}_i} = (X_{ij} : j \notin A_i)^T \) for missing covariates. Thus, equation (1) can be expressed as

\[
Y_i = X_{i,A_i}^T \beta_{A_i} + X_{i,\bar{A}_i}^T \beta_{\bar{A}_i} + \varepsilon_i,
\]

where \( \beta_{A_i} \) and \( \beta_{\bar{A}_i} \) denote the regression coefficients for the complete and incomplete covariates, respectively. For \( X_{i,\bar{A}_i} \), which is unobserved, refer to Lin et al. [9], we take the expectation of \( Y_i \) given the observed covariates. Thus, we obtain the following equation:

\[
E(Y_i | X_{i,A_i}) = X_{i,A_i}^T \beta_{A_i} + E(X_{i,\bar{A}_i}^T \beta_{\bar{A}_i} | X_{i,A_i}). \tag{2}
\]

Hence, by equation (2), we can impute the incomplete part using the information on \( X_{i,A_i} \) to obtain an estimator for \( \beta \). We use the following estimator to estimate the missing components of the covariates for unit \( i \):

\[
\hat{X}_{ij} = \frac{\sum_{i' = 1}^{n} I(A_{i'} \supset A_i \cup j) X_{i',j} K_h (X_{i',A_i} - X_{i,A_i})}{\sum_{i' = 1}^{n} I(A_{i'} \supset A_i \cup j) K_h (X_{i',A_i} - X_{i,A_i})}, \quad (j \notin A_i), \tag{3}
\]

In equation (3), \( K_h(\cdot) = K(\cdot/h)/h \), where \( K(\cdot) \) is a kernel function and \( h \) is a bandwidth. In this way, we can make use of the information as fully as possible. To tackle the problem of “the curse of dimensionality”, we further transform the estimator in equation (3) using the projective resampling method. As shown in Figure 1, for subject \( i' \), we project \( X_{i',A_i} \) onto random directions \( \{v_{b,A_i} \in \mathcal{R}^{1 \times |A_i|}, b = 1, 2, \ldots, B \} \), where \( |A| \) denotes the cardinality of a set \( A \), then obtain \( B \) kernel values using the resulting scalars \( \{X_{i',A_i}^T v_{b,A_i}, b = 1, 2, \ldots, B\} \), and finally integrate the \( B \) kernels through the geometric mean.

\[
\hat{X}_{ij} = \frac{\sum_{i' = 1}^{n} I(A_{i'} \supset A_i \cup j) X_{i',j} \left[ \prod_{b=1}^{B} K_h (X_{i',A_i}^T v_{b,A_i} - X_{i,A_i}^T v_{b,A_i}) \right]^{\frac{1}{B}}}{\sum_{i' = 1}^{n} I(A_{i'} \supset A_i \cup j) \left[ \prod_{b=1}^{B} K_h (X_{i',A_i}^T v_{b,A_i} - X_{i,A_i}^T v_{b,A_i}) \right]^{\frac{1}{B}}}, \quad (j \notin A_i). \tag{4}
\]
where $v_{b,A,i}$ is a random vector, with each entry $v_{b,A,i,j}$ chosen independently from a distribution $D$ that is symmetric about the origin with $E(v_{b,A,i,j}^2) = 1$. In practice, we usually generate $v_{b,A,i,j}$ from $N(0,1)$ or $U(-1,1)$, but random vectors from other possible distributions can also be used.

Applying the above imputation strategy, we can obtain $Z_i$ by using observed data for part $A_i$ and imputed data for part $\bar{A}_i$.

$$Z_i = (X_{ij}I(j \in A_i) + \hat{X}_{ij}I(j \notin A_i) : j = 1, 2, \cdots, p)^\top.$$

Therefore, we propose the following estimation equation:

$$U(\beta) = \frac{1}{n} \sum_{i=1}^{n} Z_i \{Y_i - Z_i^\top \beta\}. \quad (6)$$

Thus, the estimator of the regression coefficient can be solved by

$$\hat{\beta} = \left\{\sum_{i=1}^{n} Z_i Z_i^\top\right\}^{-1} \sum_{i=1}^{n} Z_i Y_i.$$

Specifically, we propose the following algorithm:
Algorithm 1 algorithm for PRIME

Input: \{(Y_i, X_i, A_i) : i = 1, 2, \cdots, n, j = 1, 2, \cdots, p\},

Output: \(\hat{\beta}\)

1: Sample entries of \(v_{b,A_i} (b = 1, 2, \cdots, B) \in \mathcal{R}^{1 \times |A_i|}\) i.i.d. from \(N(0, 1)\) or \(U(-1, 1)\)

2: for 1 \(\leq i \leq n\) and 1 \(\leq j \leq p\) do

3: 1) Apply the equation (4) to obtain \(\hat{\mathbf{X}}_{ij}\), \((j \notin A_i)\)

4: 2) Apply the equation(5) to obtain the imputed data \(Z_i\) based on \(v_{b,A_i} (b = 1, 2, \cdots, B)\)

5: Solve the closed-form equation to get \(\hat{\beta}\).

6: return \(\hat{\beta}\)

2.2 Simultaneous fitting and selection by sparse PRIME

Because the number of disease-associated biomarkers is not expected to be large, it is of great importance to take the sparse assumption into account when not all variables contribute to outcome variables. Hence, we assume the linear regression model in equation (1) is sparse and define the index set of the active and inactive predictors by \(I_1 = \{j : \beta_j \neq 0\}\) and \(I_2 = \{j : \beta_j = 0\}\), respectively. Our practical goal is to identify which biomarkers in the CSA-AKI datasets are disease-related as well as to estimate the corresponding coefficients. The main idea of sparse PRIME is to replace the estimation equation 6 with the penalized estimation equations as follows:

\[
\begin{align*}
\sum_{i=1}^{n} Z_{i1} \{Y_i - \mathbf{Z}_i^\top \mathbf{\beta}\} + \lambda_n \gamma |\beta_1|^\gamma - 1 \text{sign}(\beta_1) &= 0, \\
\vdots \\
\sum_{i=1}^{n} Z_{ip} \{Y_i - \mathbf{Z}_i^\top \mathbf{\beta}\} + \lambda_n \gamma |\beta_p|^\gamma - 1 \text{sign}(\beta_p) &= 0.
\end{align*}
\]

where \(\lambda_n \gamma |\beta_j|^\gamma - 1 \text{sign}(\beta_j)\) is the partial derivative for the penalty function with respect to \(\beta_j\).

The least absolute shrinkage and selection operator (LASSO) estimator is defined to satisfy \(\gamma = 1\). Optimizing the functions in (7) with \(\gamma = 1\) is computationally cumbersome because the functions are non-differentiable. Fortunately, the shooting algorithm proposed in Fu [16] can be used to compute the LASSO estimator. Moreover, Fu [16, 17] proved that the unique estimator of (7) is equivalent to the solution of the penalized objective function as follows:

\[
\min_{\mathbf{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - \mathbf{Z}_i^\top \mathbf{\beta}\}^2 + \frac{\lambda_n}{n} \sum_{j=1}^{p} |\beta_j|^\gamma.
\]

We can solve this penalized regression-form problem and get a sparse estimator \(\hat{\mathbf{\beta}}_s\) using the glmnet package in R.

To decrease the processing time, we can use the following sparse random projection with
i.i.d. entries:

\[ v_{b,A,i,j} = \sqrt{s_i} \begin{cases} 
1, & \text{with probability } \frac{1}{2s_i} \\
0, & \text{with probability } 1 - \frac{1}{s_i} \\
-1, & \text{with probability } \frac{1}{2s_i} 
\end{cases} \]

where Achlioptas [18] used \( s_i = 1 \) or \( s_i = 3 \) and Li et al. [19] showed that one can use \( s_i = \sqrt{|A_i|} \) or even \( s_i = \frac{|A_i|}{\log |A_i|} \) to significantly reduce the computing time with little loss of accuracy.

### 3 Theoretical properties

We study the consistency of the projection resampling least estimator \( \hat{\beta} \). Denote the true value of \( \beta \) by \( \beta_0 \). We make the following assumptions:

(A1) \( h = O(n^{-l}) \) with \( 1/4 < l < 1/2 \);

(A2) \( A_i \perp X_i \);

(A3) The kernel function \( K(\cdot) \) is a symmetric density function with compact support \([0, 1]\) and a bounded derivative;

(A4) \( \beta_0 \in B \), where \( B \) is a bounded set;

(A5) For a missing-data pattern \( A \), let \( e_{j,A}(w_A) = E(X_{ij}|X_{i,A} = w_A) \) be the conditional expectation of \( X_{ij} \) and \( f_A(w_A) \) be the density of \( X_{i,A} \). Assume \( e_{j,A}(w_A) \) and \( f_A(w_A) \) have continuous second derivatives with respect to \( w_A \) on the corresponding support;

(A6) \( E(v_{b,j}^4) < \infty \) (\( 1 \leq j \leq p \));

(A7) \( X_i \) is bounded for all \( i \geq 1 \), and the limit \( \lim_{n \to \infty} \sum_{i=1}^{n} X_i/n = X_0 \) exists;

(A8) \( D_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \to D \) and \( \frac{1}{n} \max_{1 \leq i \leq n} X_i^\top X_i \to 0 \), where \( D \) is a finite and positive definite matrix;

(A9) There exists a constant \( C_0 > 0 \), \( \inf_{w_A} P_r(A_{i'} \supset A \cup j) f_A(w_A) > C_0 \).

Assumptions (A1)–(A5) are the same as the conditions in Lin et al. [9]. Specifically, Assumption (A1) requires under-smoothing to obtain a root–\( n \) consistent estimator, which is a commonly used regularity assumption in semiparametric regression. Assumption (A2) addresses the missing data mechanism and ensures the PRIME estimator is consistent. As mentioned in Lin et al. [9], Assumption (A2) is weaker than assuming the data are missing completely at random. Assumptions (A3)–(A5) are standard in nonparametric regression. Assumption (A3) is achieved when the kernel function is the Gaussian kernel, but it is more general. Assumption (A6) is a moment bound required in Arriaga and Vempala [20] and Li et al. [19], a necessary
technical condition. Assumptions (A7) and (A8) are commonly used in penalized estimation problems [16, 17]. We assume inf_{A_p} P_r(A_r \supset \Lambda \cup j) f_A(w_A) > C_0 to make sure that there are enough samples being used to estimate e_{j,A}(w_A).

**Theorem 1.** Suppose Assumptions (A1)-(A6) hold and \( j \in I_1 = \{ j : \beta_j \neq 0 \}, j = 1, 2, \ldots, p \), as \( n \to \infty \), then \( \hat{\beta} \to \beta_0 \) in probability.

**Theorem 2.** Suppose Assumptions (A1)-(A8) hold, \( j \in I_1 = \{ j : \beta_j \neq 0 \}, j = 1, 2, \ldots, q \), \( j \in I_2 = \{ j : \beta_j = 0 \}, j = q + 1, \ldots, p \) and tuning parameter \( \lambda_n = o(\sqrt{n}) \) as \( n \to \infty \), then \( \hat{\beta}_s \to \beta_0 \) in probability.

4 Simulation

In this section, we consider several simulated scenarios to highlight the properties of PRIME in contrast to some other methods. We experimentally investigate the performance of the following methods:

**Full:** the least-squares estimator based on the full data as a benchmark;

**PRIME:** the proposed method;

**ILSE:** the iterative least-square method in Lin et al. [9];

**ML:** the maximum likelihood method proposed in Jiang et al. [4];

**CC:** the complete-case analysis method.

For each model setting with a specific choice of parameters, we repeat the simulation 100 times and evaluate the performance of models using the normalized absolute distance (NAD) and the mean squared error (MSE), defined as follows:

\[
NAD_j = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{\beta}_j - \beta_{0j}|}{\beta_{0j}}, \quad j = 1, 2, \ldots, p,
\]

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j - \beta_{0j})^2.
\]

In addition, we calculate the optimal MSE rate, defined as the proportion of times each method (except Full) produced the smallest MSE in repetitions. The MSE based on N repetitions is partitioned into MSE=Variance+Bias^2, as follows:

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j^{(i)} - \beta_{0j})^2 = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{N} \sum_{i=1}^{N} (\hat{\beta}_j^{(i)} - \hat{\beta}_j)^2 + \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j - \beta_{0j})^2,
\]

where \( \hat{\beta}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_j^{(i)} \) and \( N = 100 \) in this simulation study. For simplicity, we set bandwidth \( h = n^{-1/3} \) for PRIME, SPRIME and ILSE. In the following sections, we compare the methods using various settings for sample size, missing data rates, noise levels, and feature correlations.
Table 1: Missing pattern for all simulation examples.

| Group | Pattern | Variable |
|-------|---------|----------|
| Full  | $X_1$   | ✓        |
|       | $X_2$   | ✓        |
|       | $X_3$   | ✓        |
|       | $X_4$   | ✓        |
|       | $X_5$   | ✓        |
|       | $X_6$   | ✓        |
|       | $X_7$   | ✓        |
|       | $X_8$   | ✓        |
|       | $X_9$   | ✓        |
|       | $X_{10}$| ✓        |
|       | $X_{11}$| ✓        |
|       | $X_{12}$| ✓        |
| $A^{(1)}$ | $A_1$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_2$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_3$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_4$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_5$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_6$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
| $A^{(2)}$ | $A_7$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_8$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_9$  | ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_{10}$| ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_{11}$| ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
|         | $A_{12}$| ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ |

4.1 Scenario 1: Different noise levels

The data generation model has the linear expression

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \ i = 1, 2, \cdots, n,$$

where $n = 100, 200, p = 12, \beta = (1, -0.6, 1.5, 1, 1.2, 0.4, -1, -0.7, 1.3, 0.5, 1.1, -1.4, 0.9)^T$.

We generate $(X_{i1}, \cdots, X_{ip})$ from the multivariate normal distribution $N_p(0, \Sigma)$. We set the non-diagonal elements $\rho_{ij}$ of $\Sigma$ equal to 0.5 and the diagonal elements of $\Sigma$ equal to 1. For $\varepsilon_i$, we use the error distribution $N(0, \sigma^2)$, where $\sigma^2$ changes with $R^2 = \text{Var}(X_i^\top \beta)/\{\text{Var}(X_i^\top \beta) + \sigma^2\}$. We consider the cases in which $R^2 = 0.1, 0.2, \cdots, 0.9$.

Missing data are divided into three generative scenarios or assumptions. Missing completely at random (MCAR) means the missingness is independent of the values of the data. Missing at random (MAR) means the propensity of data to be missing depends on the observed values, whereas missing not at random (MNAR) covers the remaining scenario that the mechanism depends on the unobserved values (the variables that are missing). In our study, we consider situations that differ from the classical MCAR, MAR, and MNAR mechanism.

For each sample, variables $X_{10}, X_{11}, X_{12}$ are always available. There are twelve “typical” missing patterns considered, the details are shown in Table 1. We divide the 12 patterns into two groups. Specifically, the first group $A^{(1)}$ consists of $(A_1, A_2, \cdots, A_6)$, and the second group $A^{(2)}$ consists of the rest missing patterns. We randomly assign the missing patterns in $A^{(1)}$ to the sample with missing probability $P = a$. Furthermore, we set the missing probability of $i$th unit for the patterns in $A^{(2)}$ as $P = \{1 + \exp(b\varepsilon_i + c)\}^{-1}$. Then we randomly assign the patterns in $A^{(2)}$ to the missing samples. The settings in Table 2 are used for the missing rate (MR).
Table 2: Missing rating setting for all simulation examples.

| Missing rate | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 60%          | a   | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
|              | b   | -1.5| -1.5| -1.5| -1.5| -1.5| -2  | -2  | -4  |
|              | c   | -4  | -2  | -2  | -1.5| -1.5| -1  | -1  | -1  |
| 90%          | a   | 0.75| 0.75| 0.75| 0.75| 0.75| 0.75| 0.75| 0.65|
|              | b   | -1.5| -1.5| -2  | -2  | -3  | -3  | -3.5| -3.5| -4  |
|              | c   | -4  | -4  | -4  | -4  | -4  | -4  | -4  | -4  | -4  |

The MSE results are shown in Figures 3, 6, 9, and 12 for $R^2 = 0.2, 0.5, 0.8$. The results of optimal rate of MSE are displayed in Figures 4, 7, 10, and 13. To determine the relative performance, we rank the NADs of the five methods at each repetition. The mean NAD ranks are displayed in Figures 2, 5, 8, and 11 for $R^2 = 0.2, 0.5, 0.8$. The results of the cases not shown here are available in the supplementary materials. In general, they exhibit patterns which is similar to those shown here.

To make the comparison of Full, PRIME, and ILSE easier, the MSE bar plots for CC and ML are manually scaled because these two methods lead to much higher MSEs than the other methods. The main conclusions are as follows:

1. When $n$ and $R^2$ increase, the MSEs of Full, ILSE, and PRIME generally decrease, as expected. However, the relative performance of these three methods does not change.

2. Generally, PRIME outperforms ILSE and ML in terms of NAD, and all three significantly outperform CC. Surprisingly, PRIME was close or even superior to the Full method in estimating the coefficients of $X_{11}$ and $X_{12}$ when MR = 90%, $n = 100$, and $R^2 = 0.2$ and in estimating the coefficients of $X_{10}$ and $X_{12}$ when MR = 90%, $n = 200$, and $R^2 = 0.2$. These results confirm the superiority of the PRIME method.

3. The comparison of results shows that the proposed PRIME method performed better than the other three methods (ILSE, CC, and ML). The bias and variance decomposition figures show that ILSE produces more biased estimates than PRIME. The CC method’s estimation has extremely high variance in almost all ranges of $R^2$, and CC’s approximation error was larger when $R^2$ was not very high. Furthermore, the CC method produces biased estimates, as expected, because of the missing-data mechanism. The ML performance is not stable: ML’s performance is close to our proposed method’s when $n = 200$, MR = 90%, and $R^2 = 0.8$, but ML has the poorest performance when $n = 100$, MR = 90%, and $R^2 = 0.8$ because ML estimators can be highly biased when the MAR assumption does not hold.

4. The optimal rates of MSE show a proportion over 40% for PRIME, and this indicates that PRIME yields the smallest MSE of all the competitors in more than 40% of the trials in Scenario 1. When PRIME yields the optimal rate, ILSE or ML most often yields the second-smallest MSE. When PRIME does not yield the lowest MSE, ILSE and ML most
often do. Not surprisingly, CC rarely produces the smallest MSE except when MR = 60% and $R^2 = 0.9$.

Figure 2: NAD with $n = 100$ and 90% missing data for different methods.

Figure 3: MSE with $n = 100$ and 90% missing data for different methods.
Figure 4: Optimal rate of MSE with \( n = 100 \) and 90\% missing data for different methods.

Figure 5: NAD with \( n = 200 \) and 90\% missing data for different methods.
Figure 6: MSE with $n = 200$ and 90% missing data for different methods.

Figure 7: Optimal rate of MSE with $n = 200$ and 90% missing data for different methods.
Figure 8: NAD with $n = 100$ and 60% missing data for different methods.

Figure 9: MSE with $n = 100$ and 60% missing data for different methods.
Figure 10: Optimal rate of MSE with $n = 100$ and 60% missing data for different methods.

Figure 11: NAD with $n = 200$ and 60% missing data for different methods.
4.2 Scenario 2: Varying correlation between variables

To compare the methods with different correlations, we consider the correlation between $X_i$ and $X_j$ in four situations: $p_1$ for $\rho_{ij} = 0.2$, $p_2$ for $\rho_{ij} = 0.5$, $p_3$ for $\rho_{ij} = 0.8$ and $p_4$ for $\rho_{ij} = 0.8|j-i|$. Here, we set $\sigma^2$ with $R^2 = 0.7$. All other aspects remain the same as in Scenario 1. For the missing rate, two settings are considered, where

- $(a, b, c) = (0.1, -2, -1)$, so that the missing rate is approximately 60%,
- $(a, b, c) = (0.7, -3.5, -4)$, so that the missing rate is approximately 90%.
The NAD and MSE results are shown in Figures 14, 15, 17, 18, 20, 21, 23, and 24. The optimal MSE rates are shown in Figures 16, 19, 22, and 25. The main conclusions are as follows:

1. Scenario 2 results yield conclusions similar to those of Scenario 1. PRIME produces the smallest NADs and MSEs in almost all cases. As shown in Figures 21 and 24, although CC has the smallest estimation error, its estimation variance is extremely high, which causes problems in the MSE.

2. The optimal rate results show that PRIME has obvious advantages over other methods because it produces the smallest MSE in almost all situations except when MR = 90%, n = 200, and $\rho_{ij} = 0.8$. However, when $\rho_{ij}$ increases, the gap between PRIME and other methods increases. CC still has the worst MSE performance among the four methods.

Figure 14: NAD with $n = 100$ and 90% missing data for different methods.

Figure 15: MSE with $n = 100$ and 90% missing data for different methods.
Figure 16: Optimal rate of MSE with $n = 100$ and 90% missing data for different methods.

Figure 17: NAD with $n = 200$ and 90% missing data for different methods.
Figure 18: MSE with $n = 200$ and 90% missing data for different methods.

Figure 19: Optimal rate of MSE with $n = 200$ and 90% missing data for different methods.
Figure 20: NAD with $n = 100$ and 60% missing data for different methods.

Figure 21: MSE with $n = 100$ and 60% missing data for different methods.
Figure 22: Optimal rate of MSE with $n = 100$ and 60% missing data for different methods.

Figure 23: NAD with $n = 200$ and 60% missing data for different methods.
4.3 Scenario 3: Taking sparse structure into consideration

In this scenario, we illustrate the proposed SPRIME by studying the data from simulations. We consider penalized Full (denoted as SFull), complete-case analysis with a penalty (denoted as SCC), ILSE, and ML as the alternatives. ILSE in Lin et al. [9] and ML in Jiang et al. [4] were proposed without considering the sparse assumption; hence, we use them directly instead of using the penalized estimation form. We acknowledge that there are other approaches such as those in Xue and Qu [21] that can be used to address a high-dimensional missing-data problem. However, the missing-data patterns in these methods are different from the individual-specific case.
The model used to generate data has the linear expression

\[ Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \]

where \( n = 200, p = 30, (\beta_1, \cdots, \beta_{12})^\top = (1, -0.6, 1.5, 1, 1.2, 0.4, -1, -0.7, 1.3, 0.5, 1.1, -1.4, 0.9), \) and \( \beta_j = 0 \) \((j = 13, \cdots, 30)\). We generate \((X_{i1}, \cdots, X_{ip})\) from the multivariate normal distribution \( N_p(0, \Sigma) \). We set the non-diagonal element \( \rho_{ij} \) of \( \Sigma \) equal to 0.5. For \( \varepsilon_i \), we use the error distribution \( N(0, \sigma^2) \), where \( \sigma^2 \) changes with \( R^2 = \text{Var}(X_i^\top \beta)/\{\text{Var}(X_i^\top \beta) + \sigma^2\} \). We consider only the case in which \( R^2 = 0.7 \).

When taking sparse structure into consideration, the coefficients \( \beta_j \) \((j = 13, \cdots, 30)\) are equal to 0. Thus, the criterion, NAD used in Scenario 1 and 2, is no longer meaningful. So we only use MSE and the optimal rate of MSE to assess the performance of different methods. Several conclusions can be drawn from the figures:

1. Scenario 3 results yield conclusions similar to those of Scenarios 1 and 2. From Figure 26 we can see that SPRIME produces the smallest MSEs in almost all cases.

2. When MR=60%, the optimal rates of MSE are 0.95, 0.01, 0.00, 0.04 for SPRIME, ILSE, CC and ML, respectively. When MR=90%, the optimal rates of MSE are 0.94, 0.05, 0.01, 0.00 for SPRIME, ILSE, CC and ML, respectively. Not surprisingly, SPRIME that considers the sparse structure performs better than ILSE and ML without the penalty, which reconfirms the superiority of PRIME-type approach.

5 Cardiac surgery-associated acute kidney injury study

In this example, we illustrate the proposed method by analyzing data regarding Cardiac surgery-associated acute kidney injury (CSA-AKI). CSA-AKI is the second condition for acute kidney
injury in the intensive care setting and sometimes causes death [2]. However, because of a
general lack of effective treatment for CSA-AKI, tools or methods for earlier identification are
very important for prevention and management of the syndrome. To find more predictive
biomarkers for CSA-AKI, Chen et al. [2] collected 32 plasma cytokines, including CTACK,
FGFa, G-CSF, HGF, interferon-α2, interferon-gamma (IFN-γ), IL-1α, IL-1β, IL-2, IL-4, IL-
6, IL-7, IL-8, IL-9, IL-10, IL-12p70, IL-12p40, IL-16, IL-17α, IL-18, IP-10, MCP-1, MCP-3,
M-CSF, MIF, MIG, MIP-1α (macrophage inflammatory protein-1 α), MIP-1β (macrophage
inflammatory protein-1 β), SCF, SCGF-β, SDF-1α, and tumor necrosis factor-α. CSA-AKI
severity is evaluated primarily by deltaScore, which is measured by serum creatinine alterations
before and after surgery. Serum creatinine concentrations before and after surgery are measured
by an identical testing platform in the clinical laboratory of the hospital.

We use the continuous-variable deltaScore as the response. For simplicity, we conduct
a standardized transformation to scale the both response and covariates. Furthermore, we
exclude subjects with missing deltaScores because the aforementioned methods (PRIME and
SPRIME) apply primarily to the missing covariates. Finally, 321 patients are enrolled for
statistical analysis, of which only approximately 60% have complete covariate information.

Because the number of related variables is not expected to be large, as in Scenario 3 in
the simulation study, we use SPRIME and SCC to simultaneously select and estimate the
coefficients of the factors that might shed light on the deltaScore. We also use ILSE and
ML directly without considering the sparse assumption. The regression coefficient estimates
obtained from the four methods are listed in Table 3. Among them, IL-8, IL-10, IFN-γ, IL-16,
and MIP-α are also found to be related to CSA-AKI in Chen et al. [2]. However, in real-world
data, it is difficult to objectively evaluate the performance of candidate methods. Therefore,
we delete the subjects with missing covariates and construct missing data manually for the
complete-case data of CSA-AKI. For the same reason as before in Scenario 3, we consider only
MSE to evaluate the estimation accuracy. However, because of the unknown true coefficients,
we are unable to evaluate MSE as described in the simulation. Hence, we calculate MSE_{Full}
instead, as follows:

$$\text{MSE}_{\text{Full}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_{\text{Full},j} - \hat{\beta}_j)^2,$$

where $\hat{\beta}_{\text{Full},j}$ ($j = 1, 2, \ldots, p$) are the estimated coefficients obtained using Full method.

The setting of missing-data patterns is the same as that in the simulation study except
for the missing probability function. We randomly assign the missing patterns in $A^{(1)}$ to the
sample with missing probability $P = a$. Furthermore, we set the missing probability of the ith unit
for the patterns in $A^{(2)}$ as $P = \{1 + \exp(b\epsilon_i + c)\}^{-1}$, where $\epsilon_i = Y_i - X_i^T \hat{\beta}_{\text{Full}}$. Then,
we randomly assign the patterns in $A^{(2)}$ to the missing samples. For the missing rate, we
set $(a, b, c) = (-1.5, -2, 0.4)$. Consequently, the missing rate is approximately 90%. This is
repeats $N = 100$ times to randomly generate missing data. Here, like in simulation, we also set
bandwidth $h = n^{-1/3}$.

The MSE_{Full} results are shown in Figure 27. The optimal rates of MSE are 1.00, 0.00, 0.00,
0.00 for SPRIME, ILSE, CC and ML, respectively. The results show that SPRIME has advan-
tages over the other methods in estimation accuracy as it produces the smallest MSE_{Full}. The
missing mechanism and the wrong model assumption may give rise to the worse performance of ILSE, CC and ML.

![Real data chart]

Figure 27: MSE with real data for different methods.

6 Conclusions

In this study, we propose a projective resampling imputation mean estimation method to estimate the regression coefficients for a high rate of missing-data covariates. Our first set of random features projects data points onto a randomly chosen line and then averages the resulting scalar values to yield a comprehensive result. The random lines are drawn from the standard normal distribution to ensure less loss of information. We experimentally evaluate the performance of the PRIME method, and the results show that the proposed method is a feasible alternative.

However, the aforementioned work has been developed for a classical setting. Developing PRIME to combine generalized linear models or Cox models with missing data warrants future research. Furthermore, we considered only missing covariates even though it is common to encounter cases where both covariates and responses are missing. Hence, developing methods to address practical issues will be the focus of our future work.
Table 3: Regression coefficients of CPLSD with regression coefficient estimates (The symbol "-" means that the current variable or predictor has not been selected).

| Variable    | SPRIME | SCC    | ILSE    | ML     |
|-------------|--------|--------|---------|--------|
| age         | -0.058 | -0.099 | -0.167  | -0.012 |
| BMI         | -      | -      | 0.040   | 0.048  |
| hospitalized time | 0.051 | 0.045 | 0.069   | 0.080  |
| CTACK       | -      | 0.020  | 0.207   | 0.117  |
| FGFα        | -      | -      | 0.137   | 0.137  |
| G-CSF       | -      | -      | -0.403  | -0.462 |
| HGF         | -      | -      | -0.120  | -0.162 |
| IFN-α2      | -      | -      | 0.233   | 0.393  |
| IFN-γ       | 0.089  | 0.063  | 0.207   | 0.199  |
| IL-1α       | -      | -      | -0.126  | -0.113 |
| IL-1β       | -      | -      | -0.257  | -0.256 |
| IL-2        | -      | -      | 0.349   | 0.280  |
| IL-4        | -      | -      | -0.011  | -0.012 |
| IL-6        | -      | -      | -0.126  | -0.148 |
| IL-7        | -      | -      | -0.165  | -0.182 |
| IL-8        | 0.149  | 0.155  | 0.527   | 0.624  |
| IL-9        | -      | -0.008 | -0.020  | -0.034 |
| IL-10       | 0.040  | 0.023  | 0.043   | 0.064  |
| IL-12p70    | -      | -      | 0.347   | 0.361  |
| IL-12p40.   | -      | -      | -0.029  | -0.099 |
| IL-16       | 0.100  | 0.109  | 0.143   | 0.118  |
| IL-17α      | -      | -      | -0.322  | -0.303 |
| IL-18       | -      | -      | -0.071  | -0.080 |
| IP-10       | -      | -      | -0.057  | -0.170 |
| MCP-1       | -      | -      | 0.081   | 0.060  |
| MCP-3       | -      | -      | -0.099  | -0.233 |
| M-CSF       | -      | 0.080  | 0.024   | 0.042  |
| MIF         | -      | -      | -0.131  | -0.146 |
| MIG         | 0.024  | 0.009  | 0.169   | 0.342  |
| MIP-1α      | 0.036  | 0.039  | 0.315   | 0.330  |
| MIP-1β      | -      | -      | -0.187  | -0.165 |
| SCF         | 0.175  | 0.143  | 0.166   | 0.175  |
| SCGF-β      | 0.093  | 0.0923 | 0.060   | 0.121  |
| SDF-1α      | -      | -      | -0.096  | -0.057 |
| preLVEF     | -      | -      | 0.056   | 0.073  |
| TNF-α       | -      | -      | -0.100  | -0.071 |
References

[1] Mengyun Wu, Jian Huang, and Shuangge Ma. Identifying gene-gene interactions using penalized tensor regression. *Statistics in medicine*, 37(4):598–610, 2018.

[2] Zhongli Chen, Liang Chen, Guangyu Yao, Wenbo Yang, Ke Yang, and Chenglong Xiong. Novel blood cytokine-based model for predicting severe acute kidney injury and poor outcomes after cardiac surgery. *Journal of the American Heart Association*, 9(22):e018004, 2020.

[3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–22, 1977.

[4] Wei Jiang, Julie Josse, Marc Lavielle, and TraumaBase Group. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics and Data Analysis*, 2020.

[5] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013.

[6] BaoLuo Sun and Eric J. Tchetgen Tchetgen. On inverse probability weighting for non-monotone missing at random data. *Journal of the American Statistical Association*, 113(521):369–379, 2018.

[7] Judith M Conn, Kung-Jong Lui, and Daniel L McGee. A model-based approach to the imputation of missing data: Home injury incidences. *Statistics in Medicine*, 8(3):263–266, 1989.

[8] Xiaoyan Yin, Daniel Levy, Christine Willinger, Aram Adourian, and Martin G Larson. Multiple imputation and analysis for high-dimensional incomplete proteomics data. *Statistics in medicine*, 35(8):1315–1326, 2016.

[9] Huazhen Lin, Wei Liu, and Wei Lan. Regression analysis with individual-specific patterns of missing covariates. *Journal of Business and Economic Statistics*, 19(3):231–253, 2019.

[10] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[11] Leonard J. Schulman. Clustering for edge-cost minimization (extended abstract). *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 547–555, 2000.

[12] D.L. Donoho. Compressed sensing. *IEEE Transactions Information Theory*, 52:1289–1306, 2006.

[13] Qinfeng Shi, Hanxi Li, and Chunhua Shen. Rapid face recognition using hashing. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2753–2760, 2010.
[14] Odalric-Ambrym Maillard and Rémi Munos. Linear regression with random projections. *Journal of Machine Learning Research*, 13(1):2735–2772, 2012.

[15] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: approximating kernel expansions in loglinear time. *ICML’13 Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, 2013.

[16] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[17] Wenjiang J Fu. Penalized estimating equations. *Biometrics*, 59(1):126–132, 2003.

[18] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[19] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.

[20] Rosa I Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 16:161–182, 2006.

[21] Fei Xue and Annie Qu. Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, pages 1–14, 2020.

[22] Kani Chen, Shaojun Guo, Liuquan Sun, and Jane-Ling Wang. Global partial likelihood for nonparametric proportional hazards models. *Journal of the American Statistical Association*, 105(490):750–760, 2010.
Appendix

Proof of Theorem 1

Proof. Following the proof of Lin et al. [9], we first define

\[
\hat{E}_{j,A}(w_A) = \frac{\sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) X_{i'j} \prod_{b = 1}^{B} \left[ K_h \left( X_{i',A}^T v_{b,A} - w_{A}^T v_{b,A} \right) \right] \frac{1}{\pi}}{\sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) \prod_{b = 1}^{B} \left[ K_h \left( X_{i',A}^T v_{b,A} - w_{A}^T v_{b,A} \right) \right] \frac{1}{\pi}}, \quad (j \notin A)
\]

and then

\[
Z_i = (X_{ij} I(j \in A_i) + \hat{E}_{j,A_i}(X_{i,A_i}) I(j \notin A_i : 1 \leq j \leq p),
\]

\[
U(\beta) = \frac{1}{n} \sum_{i = 1}^{n} Z_i \{Y_i - Z_i^\top \beta\}.
\]

Let

\[
z_i = (X_{ij} I(j \in A_i) + e_{j,A_i}(X_{i,A_i}) I(j \notin A_i : 1 \leq j \leq p))^\top
\]

\[
u(\beta) = E \left[ z_i(\beta) \{X_i^\top \beta_0 - z_i^\top \beta\} \right].
\]

Similar to Lin et al. [9], we first prove that \(\beta_0\) is a solution of \(u(\beta)\). We have

\[
E(Y_i|X_{i,A_i}, A_i) = z_i^\top \beta_0.
\]

and we also have \(E(Y_i|X_{i,A_i}, A_i) = E(X_i^\top|X_{i,A_i}, A_i) \beta_0\), hence we can get \(E \left\{ (X_i - z_i^\top)|X_{i,A_i}, A_i \right\} \beta_0 = 0\). Noting that \(z_i\) is a function of \((X_{i,A_i}, A_i)\), then

\[
u(\beta_0) = E \left\{ z_i \left( X_i^\top \beta_0 - z_i^\top \beta_0 \right) \right\} = 0 \quad (8)
\]

Similar to Lin et al. [9], it is easy to show that \(\beta_0\) is the unique solution of \(u(\beta) = 0\). To prove Theorem 1, it suffices to show that

\[
\sup_{\beta \in \mathcal{B}} \|U(\beta) - u(\beta)\|_2 = 0. \quad (9)
\]

We rewrite

\[
U(\beta) - u(\beta) = U_1 - U_2 + U_3 + U_4,
\]
where

\[ U_1 = \frac{1}{n} \sum_{i=1}^{n} (Z_i - z_i) Y_i, \]
\[ U_2 = \frac{1}{n} \sum_{i=1}^{n} (Z_i Z_i^T - z_i z_i^T) \beta, \]
\[ U_3 = \frac{1}{n} \sum_{i=1}^{n} \left[ z_i \{ X_i^T \beta_0 - z_i^T \beta \} - E \left\{ z_i (X_i^T \beta_0 - z_i^T \beta) \right\} \right], \]
\[ U_4 = \frac{1}{n} \sum_{i=1}^{n} z_i \varepsilon_i. \]

Then \( \sup_{\beta \in B} \| U_3 \|_2 = o_p(1), \sup_{\beta \in B} \| U_4 \|_2 = o_p(1) \) follows from the weak law of large numbers. Noting that

\[
\hat{E}_{j,A}(w_A) - e_{j,A}(w_A) = \frac{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \{ X_{i'j} - e_{j,A}(w_A) \} \left[ \prod_{b=1}^{B} K_h \left( X_{i'j,b} A - w_A v_{i',A} \right) \right]^{\frac{1}{2}}}{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \left[ \prod_{b=1}^{B} K_h \left( X_{i'j,b} A - w_A v_{i',A} \right) \right]^{\frac{1}{2}}} \text{ (10)}
\]

For gaussian kernel generates better empirical performance than do other types of kernels, here we assume \( K_h(\cdot) \) is a gaussian kernel function, then \( \left[ \prod_{b=1}^{B} K_h \left( X_{i'j,b} A - w_{i',A} v_{b,A} \right) \right]^{\frac{1}{2}} \) can be written as exp \( \left\{ -\left\| \frac{1}{\sqrt{B}} V X_{i',A} - \frac{1}{\sqrt{B}} V w_{i',A} \right\|^2 \right\} \), where \( V \) be a \( B \times |A| \) random matrix whose entries are chosen independently from either \( N(0, 1) \) or \( U(-1, 1) \). Hence, we can show that \( \left[ \prod_{b=1}^{B} K_h \left( X_{i',A}^T v_{b,A} - X_{i',A}^T v_{b,A} \right) \right]^{\frac{1}{2}} = K_h(X_{i',A} - X_{i,A}) + o_p(1) \) using Theorem 1 in Arriaga and Vempala [20]. Then, (10) can be written as

\[
\hat{E}_{j,A}(w_A) - e_{j,A}(w_A) = \frac{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \{ X_{i'j} - e_{j,A}(w_A) \} \left[ K_h \left( X_{i',A} - w_A \right) + o_p(1) \right]}{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \left[ K_h \left( X_{i',A} - w_A \right) + o_p(1) \right]} \]
\[ = \frac{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \{ X_{i'j} - e_{j,A}(w_A) \} K_h \left( X_{i',A} - w_A \right)}{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \left[ K_h \left( X_{i',A} - w_A \right) + o_p(1) \right]} + \frac{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \{ X_{i'j} - e_{j,A}(w_A) \} o_p(1)}{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \left[ K_h \left( X_{i',A} - w_A \right) + o_p(1) \right]} = \hat{E}_1 + \hat{E}_2. \]

Under the conditions given, following the proof of theorem 1 in Lin et al. [9] and the lemma 4 in Chen et al. [22], we know that

\[
\sup_{w_A} \| E_1 \|_2 \leq \sup_{w_A} \left\| \frac{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \{ X_{i'j} - e_{j,A}(w_A) \} K_h \left( X_{i',A} - w_A \right)}{\sum_{i'=1}^{n} I (A_{i'} \supset A \cup j) \left[ K_h \left( X_{i',A} - w_A \right) + o_p(1) \right]} \right\|_2 = o_p(1)
\]
Let \( R(w_A; j) = 1/n \sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) K_h(X_{i'}, A - w_A), \) and \( \sup_{w_A} |R(w_A; j) - Pr(A_{i'} \supset A \cup j) f(w_A)| = O_p \left( \sqrt{\log n}/\sqrt{n h} + \ldots \right) \) follows from lemma 4 in Chen et al. [22]. Using Assumption (A5), then the proof of Theorem 2 is easily conducted by using the theorem 1 and 2 in Fu [16] and theorem 3 in Fu [17].

Therefore, we have

\[
\sup_{w_A} \|E_2\| \leq \sup_{w_A} \left\| \frac{1}{n} \sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) (X_{i'} - e_{j,A}(w_A)) K_h(X_{i'}, A - w_A) \right\|_2.
\]

Let \( S_{i'} = I(A_{i'} \supset A \cup j) \{X_{i'} - e_{j,A}(w_A)\} / \{C + \frac{1}{n} \sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) o_p(1)\}. \) Clearly, for the condition \( A_i \perp X_i \),

\[
E(S_{i'}) = E \left[ \frac{I(A_{i'} \supset A \cup j) X_{i'}}{C + \frac{1}{n} \sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) o_p(1)} \right] - E \left[ \frac{I(A_{i'} \supset A \cup j) E(X_{i'}|X_i, A = w_A)}{C + \frac{1}{n} \sum_{i' = 1}^{n} I(A_{i'} \supset A \cup j) o_p(1)} \right]
\]

\[
= E \left[ \frac{X_{i'}}{C + o_p(1)} \right] \left( I(A_{i'} \supset A \cup j) = 1 \right) P_r(A_{i'} \supset A \cup j)
\]

\[
- E \left\{ E \left[ \frac{X_{ij}}{C + o_p(1)} \right] X_{i', A = w_A} \right\} \left( I(A_{i'} \supset A \cup j) = 1 \right) P_r(A_{i'} \supset A \cup j)
\]

\[
= \frac{1}{C} \{ E(X_{i'}) P_r(A_{i'} \supset A \cup j) - E(E(X_{ij}|X_i, A = w_A) P_r(A_{i'} \supset A \cup j) \}
\]

\[
= \frac{1}{C} \{ E(X_{i'}) P_r(A_{i'} \supset A \cup j) - E(X_{ij}) P_r(A_{i'} \supset A \cup j) \} = 0.
\]

Then using the weak law of large number, we have

\[
\sup_{w_A} \|E_2\| \leq o_p(1).
\]

Thus, we obtain \( \sup_{\beta \in B} \|U_1\|_2 = o_p(1) \) and the same argument can also apply to \( \sup_{\beta \in B} \|U_2\|_2. \) The above convergences imply that \( \sup_{\beta \in B} \|U(\beta) - u(\beta)\|_2 = 0. \) Following the technical derivation follow from Lin et al. [9], Theorem 1 holds.

\[\square\]

**Proof of Theorem 2**

Proof. It is obvious that the equations (8) and (9) are always correct no matter the assumption \( j \in \mathcal{I}, j = 1, 2, \ldots, p \) is satisfied or not. Then, the proof of Theorem 2 is easily conducted by using the theorem 1 and 2 in Fu [16] and theorem 3 in Fu [17].

\[\square\]
Supplementary material for “Projective Resampling Imputation Mean Estimation Method for Missing Covariates Problem”

Zishu Zhan\textsuperscript{1,2}, Xiangjie Li\textsuperscript{3,*}, Jingxiao Zhang\textsuperscript{1,2,*}

June 17, 2021

\textsuperscript{1}Center for Applied Statistics, Renmin University of China,  
\textsuperscript{2}School of Statistics, Renmin University of China,  
\textsuperscript{3}State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Abstract

In this supplementary material, we provide additional numerical results that support our conclusions.

\textbf{S1. Additional simulation results for Scenario 1}

![Graphs](image.png)

Figure 1: NAD with $n = 100$ and 90\% missing data for different methods.
Figure 2: NAD with $n = 100$ and 90% missing data for different methods.

Figure 3: MSE with $n = 100$ and 90% missing data for different methods.

Figure 4: MSE with $n = 100$ and 90% missing data for different methods.
Figure 5: NAD with $n = 200$ and 90% missing data for different methods.

Figure 6: NAD with $n = 200$ and 90% missing data for different methods.

Figure 7: MSE with $n = 200$ and 90% missing data for different methods.
Figure 8: MSE with $n = 200$ and 90% missing data for different methods.

Figure 9: NAD with $n = 100$ and 60% missing data for different methods.

Figure 10: NAD with $n = 100$ and 60% missing data for different methods.
Figure 11: MSE with $n = 100$ and 60% missing data for different methods.

Figure 12: MSE with $n = 100$ and 60% missing data for different methods.

Figure 13: NAD with $n = 200$ and 60% missing data for different methods.
Figure 14: NAD with \( n = 200 \) and 60% missing data for different methods.

Figure 15: MSE with \( n = 200 \) and 60% missing data for different methods.

Figure 16: MSE with \( n = 200 \) and 60% missing data for different methods.