Statistica Sinica **35** (2025), 431-456
doi:https://doi.org/10.5705/ss.202022.0104

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL LINEAR REGRESSION WITH BLOCKWISE MISSING DATA

Fei Xue, Rong Ma and Hongzhe Li*

Purdue University, Harvard University and University of Pennsylvania

Abstract: Blockwise missing data occur frequently when we integrate multisource or multimodality data, in which different sources or modalities contain complementary information. In this study, we consider a high-dimensional linear regression model with blockwise missing covariates and a partially observed response variable. Under this framework, we propose a computationally efficient estimator for the regression coefficient vector based on carefully constructed unbiased estimating equations and a blockwise imputation procedure, and obtain its rate of convergence. Furthermore, building on an innovative projected estimating equation technique that intrinsically corrects any bias in the initial estimator, we propose a nearly unbiased estimator for each individual regression coefficient, which is asymptotically normally distributed under mild conditions. Based on these debiased estimators, we construct asymptotically valid confidence intervals and statistical tests for each regression coefficient. The results of our numerical studies and an application to data from the Alzheimer's Disease Neuroimaging Initiative show that the proposed method outperforms existing methods, and benefits more from unsupervised samples than existing methods do.

Key words and phrases: Blockwise imputation, data integration, projected estimating equation.

1. Introduction

The problem of blockwise missing data arises when we integrate data from multiple modalities, sources, or studies. For instance, the Alzheimer's Disease Neuroimaging Initiative (ADNI) study collects data from magnetic resonance imaging (MRI), positron emission tomography (PET) imaging, genetics, cerebrospinal fluid, cognitive tests, and demographic information of patients (Mueller et al., 2005). However, because some subjects do not have MRI or PET images, the biomarkers related to the images can be completely missing for these subjects. As a result, when we integrate data from multiple sources, and group patients based on their missing patterns, blocks of values may be missing, as illustrated in Figure 1(a), where white areas represent the missing blocks. Multimodality data also appear in modern genomic studies of complex diseases. For example, the Genotype-Tissue Expression (GTEx) study has collected RNA-seq gene expression data from over 45 tissues of more than 800 donors

*Corresponding author. E-mail: hongzhe@upenn.edu

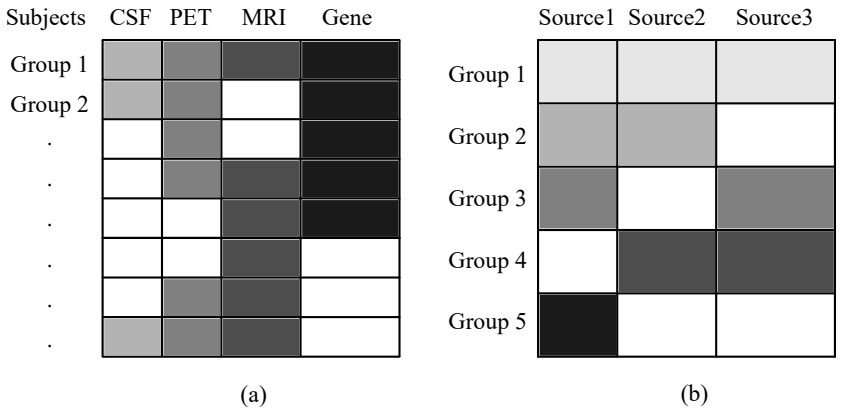

Figure 1. White areas represent missing blocks, while shaded areas represent observed blocks. (a) Missing structure for ADNI data. (b) A blockwise missing example.

(Lonsdale et al., 2013). In this case, the gene expression data in the GTEx study are blockwise missing if a tissue sample is not available.

Many important scientific questions can be answered by using an association or regression analysis. In this case, for data sets with blockwise missing covariates, the response variable is often also partially missing across the samples, for example, this situation could occur when the outcomes are expensive to collect, such as in electronic health records databases, where labeling the outcome for each individual is costly and time consuming (Kohane, 2011). In the GTEx study, samples are collected only from non-diseased tissue samples across individuals (GTEx Consortium, 2017), implying that the response is only partially observed when we predict a gene expression in one tissue using gene expression levels in other tissues.

Therefore, to make the most use of such data sets, it is essential to develop methods that are adaptive and can effectively use extra unsupervised samples to infer the underlying models.

In this study, we consider a linear regression model

$$\mathcal{Y} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{\beta} + \epsilon, \tag{1.1}$$

where $\mathcal{Y}$ is the response variable, $\boldsymbol{\mathcal{X}}$ is a p-dimensional random vector of regression covariates, $\boldsymbol{\beta}$ is a p-dimensional regression sparse coefficient vector, and ϵ is a centered sub-Gaussian random variable with variance σ^2 and independent of $\boldsymbol{\mathcal{X}}$. Let s be the number of relevant covariates with nonzero coefficients. Suppose that $\boldsymbol{\mathcal{X}}$ consists of covariates from S data sources. For instance, there are four sources in Figure 1(a), and three sources in Figure 1(b). We further suppose that all samples are drawn independently from $(\boldsymbol{\mathcal{X}}, \mathcal{Y})$ in (1.1) before going through certain missingness mechanisms.

Throughout, we allow the response variable to be missing. Specifically, we let the index set of all samples be $\mathcal{D} = \{1, \ldots, N+n\} = \mathcal{D}_1 \cup \mathcal{D}_2$, where $\mathcal{D}_1$ is the index set of the samples for which the response variable is not observable, $\mathcal{D}_2$ is the index set of the samples with observed responses, and N and n are the numbers of samples in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. For simplicity, we slightly abuse the terminology, and refer to the samples in $\mathcal{D}_1$ as the "unsupervised samples," and refer to the samples in $\mathcal{D}_2$ as the "supervised samples." We let $\boldsymbol{y}$ denote the $(N+n)$-dimensional vector consisting of all samples of the response, and let $\boldsymbol{X}$ denote the $(N+n) \times p$ design matrix, where $\boldsymbol{y}$ and $\boldsymbol{X}$ can both have missing values. In Section S1 of the Supplementary Material, we provide a table explaining all notation used in this paper.

We assume that the covariates are blockwise missing. Specifically, we assume that there are R groups of samples in $\mathcal{D}$, with the same missing covariate indices within each group, and the missing covariates consist of variables in one or several data sources. This gives rise to missing blocks in the design matrix, as shown in Figure 1. There are $R = 8$ missing groups in Figure 1(a), and $R = 5$ missing groups in Figure 1(b). For any $i = 1, \ldots, N+n$, we let ξ_i be the group label of the ith sample, which takes random values in $\{1, \ldots, R\}$. For any $r = 1, \ldots, R$, we let $\mathcal{S}(r) \subseteq \mathcal{D}$ be the index set of the samples in Group r. Our goal is to study statistical inference for the high-dimensional regression vector $\boldsymbol{\beta}$ in (1.1) based on such partially observed responses and blockwise missing covariates.

In general, there are three types of missingness mechanisms (Little and Rubin, 2019). If the missingness of a missing variable is independent of the values of both the missing variables and the observed variables, then we refer to this as missing completely at random (MCAR). If the missingness can be fully accounted for by observed variables for which we have complete information, then the missing mechanism is missing at random (MAR). If the missingness depends on the values of the missing variables, then the missing mechanism is called missing not at random (MNAR). For blockwise missing covariates, the corresponding missing mechanism depends on the relationship between $\xi_i (1 \le i \le N)$ and the covariates. For example, if ξ_i depends only on covariates observed in all groups, then the missingness mechanism of the blockwise missing covariates is MAR. We focus mainly on MAR, but do also investigate MNAR in simulations in Section 5.

1.1. Related works

Several methods have been developed recently related to blockwise missing data (Yuan et al., 2012; Xiang et al., 2014; Yu et al., 2020; Cai, Cai and Zhang, 2016; Xue and Qu, 2021). In particular, Yuan et al. (2012) studied the integration of large-scale brain imaging data sets from multiple imaging modalities, where data are blockwise missing, because each modality contains missing measurements. They propose dividing the blockwise missing data into

several learning tasks, based on the availability of the data sources, and use penalization to encourage the selection of a common set of features across all tasks. Xiang et al. (2014) extended the method by letting feature-level parameters be the same across all tasks, which is beneficial for the prediction of subjects with new missing patterns. Moreover, they included parameters for source-level weights to reflect the effectiveness of each source. Nevertheless, none of these existing methods aim to construct confidence intervals or test hypotheses for the regression models, nor do they incorporate a partially observed response with blockwise missing data.

In general, the simplest approach to handle missing data is to restrict the analysis to complete cases. However, this might induce bias if the missingness mechanism is not completely at random. The inverse probability weighting (IPW) is widely used to correct this bias (Little and Rubin, 2019) by modeling the probability of being a complete case, given some predictors, and then reweighting complete cases using the inverse of the estimated probability. The augmented IPW methods improve the IPW by combining it with imputation of missing values (Robins, Rotnitzky and Zhao, 1994; Qin, Zhang and Leung, 2017; Seaman and Vansteelandt, 2018). However, these methods are not directly applicable or easily extendable to blockwise missing data, without sacrificing efficiency. This is because IPW-related methods usually only consider whether or not a subject is completely observed, and cannot fully use the blockwise missing structure of the blockwise missing covariates.

With regard to statistical inference for high-dimensional regression models under fully observed settings, several studies employ a bias correction of regularized estimators, including thos of Javanmard and Montanari (2014), van de Geer et al. (2014), Zhang and Zhang (2014), Ning and Liu (2017), Javanmard and Montanari (2018), and Neykov et al. (2018), among many others.

More recently, high-dimensional inference problems with partially observed responses have been studied (Bellec et al., 2018; Zhang and Bradic, 2019; Cai and Guo, 2020; Deng et al., 2020). However, none of these methods address the problem of missing covariates. In particular, to the best of our knowledge, there is no existing method that focuses on statistical inference for a high-dimensional regression with blockwise missing data.

1.2. Main contributions

In this study, we build on a blockwise imputation (BI) procedure and carefully constructed unbiased estimating equations to account for structural missing covariates and a partially observed response variable. As such, we propose a computationally efficient sparse estimator for a high-dimensional regression coefficient vector, and obtain its theoretical properties under mild regularity conditions. Importantly, unlike most existing methods, our method does not require fully observed samples in the data, and benefit automatically from

additional unsupervised samples, until achieving the optimal rate of convergence of fully observed samples.

In addition, we develop an innovative projected estimating equation technique that leverages all available data, including the unsupervised samples, to correct the bias in the initial sparse estimator, and to obtain nearly unbiased estimators for the individual regression coefficients. These estimators are shown to be asymptotically normally distributed, with a variance that is minimized by construction. By carefully analyzing these debiased estimators, we can construct asymptotically valid confidence intervals and statistical tests about each regression coefficient accordingly. In particular, our theoretical analysis provides important insights about the benefits of using unsupervised samples on the proposed inference procedures, revealing their important role in constructing estimators with competitive efficiency (see also the discussions after Theorems 1 and 2).

1.3. Notation

Throughout, for a vector $\boldsymbol{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$, we define the ℓ_p-norm $\|\boldsymbol{a}\|_p = \big(\sum_{i=1}^n a_i^p\big)^{1/p}$, the ℓ_0-norm $\|\boldsymbol{a}\|_0 = \sum_{i=1}^n 1\{a_i \neq 0\}$, and the ℓ_∞-norm $\|\boldsymbol{a}\|_\infty = \max_{1\le j\le n} |a_i|$. For an index set $\mathcal{E} \subset \{1, \ldots, n\}$, we denote $\boldsymbol{a}_\mathcal{E}$ as the subvector of $\boldsymbol{a}$ consisting of all the components a_j, where $j \in \mathcal{E}$. In addition, $\boldsymbol{a}_{-j} \in \mathbb{R}^{n-1}$ denotes the subvector of $\boldsymbol{a}$ without the jth component. For a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times q}$, $\lambda_i(\boldsymbol{A})$ denotes the ith largest singular value of $\boldsymbol{A}$, and $\lambda_{\max}(\boldsymbol{A}) = \lambda_1(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A}) = \lambda_{\min(p,q)}(\boldsymbol{A})$. For index sets $S_1 \subseteq [1:p]$ and $S_2 \subseteq [1:q]$, we denote $\boldsymbol{A}_{S_1S_2}$ as the submatrix of $\boldsymbol{A}$ consisting of its entries in the rows indexed by S_1 and the columns indexed by S_2. We denote $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{ij}|$. For any positive integer n, we denote the set $\{1, 2, \ldots, n\}$ as $[1:n]$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$, $a_n \ll b_n$, or $b_n \gg a_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$, or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \le Cb_n$, for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For a set A, we denote $|A|$ as its cardinality.

2. Parameter Estimation using Blockwise Imputation

2.1. Blockwise imputation (BI)

The BI procedure is able to use more information from incomplete samples (or cases) than traditional single regression imputation (SI) methods do, which impute missing values via regression models using all the observed variables as the predictors (Baraldi and Enders, 2010; Zhang, 2016; Campos et al., 2015). For example, in Figure 1(b), the traditional SI method imputes missing values in Group 2 by modeling the relationship between the variables in Source 3 and all other variables. This relationship can be estimated based on complete samples

in Group 1. However, Groups 3 and 4 also contain information about Source 3 variables, but are not used by the SI. In contrast, the BI imputes the missing values in a group by using both the dependence between the missing variables and all the observed variables in this group, and the dependence between the missing variables and part of the observed variables, which could lead to several imputations for each missing value. The additional imputations based on part of the observed variables incorporate information from incomplete groups, that is, Groups 3 and 4 in Figure 1(b), because these incomplete groups can be used to estimate the latter dependence.

Specifically, for each missing group, the first step in BI is to determine the groups that can be used to construct an association between the missing variables and at least part of the observed variables in this group. For each Group $r \in [1 : R]$, we let $\mathcal{G}(r) \subseteq [1 : R]$ be the index set of the groups in which all the missing variables of Group r and the variables in at least one of the other sources are observed, and let $a(r), a(r)^c \subseteq [1 : p]$ be the index sets of the observed variables and missing variables, respectively, in Group r. For example, when there are three sources of data with $R = 5$ missing groups, as shown in Figure 1(b), then $\mathcal{G}(2) = \{1, 3, 4\}$, and $a(2)^c$ consists of indices of the covariates in Source 3. Group 5 is not in $\mathcal{G}(2)$, because it does not contain any information about the variables in Source 3 that are missing in Group 2. If Group r is completely observed, that is, there are no missing values in Group r, we let $\mathcal{G}(r) = \{r\}$.

In this paper, we assume, without loss of generality, that $|\mathcal{G}(r)| \geq 1$, for each $r \in [1 : R]$, implying that each covariate is observed in at least one group. This assumption is equivalent to that, for each missing variable in Group r, there is at least one group of samples reflecting the association between this missing variable and at least part of the observed variables in Group r. Note that this assumption does not require the existence of complete samples, because incomplete groups could also contain values for both missing variables and some observed variables in Group r.

In the second step of BI, we impute missing values in Group r based on each of the groups in $\mathcal{G}(r)$. Specifically, for any sample i in Group $r \in [1 : R]$ (i.e., $i \in \mathcal{S}(r)$), if the variable X_{ij} is missing ($j \in a(r)^c$), then for any Group $k \in \mathcal{G}(r)$, we impute X_{ij} by $E(X_{ij}|\boldsymbol{X}_{ia(r,k)})$, where X_{ij} is the (i, j) element in the design matrix $\boldsymbol{X}$, and $a(r, k) \subseteq [1 : p]$ is an index set of covariates observed in both Groups r and k. Throughout, for each $r \in [1 : R]$ and $i \in \mathcal{S}(r)$, we define $\boldsymbol{X}_i^{(k)} = (X_{i1}^{(k)}, \ldots, X_{ip}^{(k)})^\top$ as the imputed random vector for sample i according to Group $k \in \mathcal{G}(r)$, so that $X_{ij}^{(k)} = E(X_{ij}|\boldsymbol{X}_{ia(r,k)})$ if the jth covariate X_{ij} is missing in the ith sample $\boldsymbol{X}_i$, otherwise $X_{ij}^{(k)} = X_{ij}$. Note that the superscript (k) indicates the conditional expectation imputation based on Group k.

Often, we can estimate the conditional expectation $E(X_{ij}|\boldsymbol{X}_{ia(r,k)})$ by fitting a linear regression model between X_{ij} and the random vector $\boldsymbol{X}_{ia(r,k)}$ using the samples in Group k. To account for high dimensionality, we consider the Dantzig

selector (Candes and Tao, 2007), defined as

$$\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} = \underset{\gamma\in\mathbb{R}^{|a(r,k)|}}{\operatorname{argmin}} \|\boldsymbol{\gamma}\|_1, \quad \text{subject to } \left\|\boldsymbol{X}_{\mathcal{S}(k)j} - \boldsymbol{X}_{\mathcal{S}(k)a(r,k)}\boldsymbol{\gamma}\right\|_\infty \le \tau, \tag{2.1}$$

where $\tau > 0$ is a tuning parameter. Then, we can approximate the imputed variable $X_{ij}^{(k)} = E(X_{ij}|\boldsymbol{X}_{ia(r,k)})$ by $\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$. The imputed values are deterministic, given the data, and may be biased in the high-dimensional setting. Below, we carefully analyze such an imputation error (Section 3), and propose a bias-correction procedure to construct asymptotically unbiased estimators for the components of $\boldsymbol{\beta}$ (Section 2.3).

For each $r \in [1:R]$ and $i \in \mathcal{S}(r)$, we define $\widehat{\boldsymbol{X}}_i^{(k)} = (\widehat{X}_{i1}^{(k)}, \ldots, \widehat{X}_{ip}^{(k)})^\top$ as the actual imputed observations of sample i based on Group $k \in \mathcal{G}(r)$, where $\widehat{X}_{ij}^{(k)} = \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$ if the jth covariate is missing in the ith sample $\boldsymbol{X}_i$, otherwise $\widehat{X}_{ij}^{(k)} = X_{ij}$. Importantly, because for each group r, $\mathcal{G}(r)$ could contain multiple elements (e.g., $|\mathcal{G}(2)| = 3$ in Figure 1(b)), there could be multiple imputations for the missing blocks in this group, each associated with a distinct $k \in \mathcal{G}(r)$. Finally, we obtain the theoretical value for the tuning parameter τ in (2.1) in Section 3; in practice, τ can be determined using cross-validation (Section 5).

2.2. Construction of estimating equations and the proposed estimator

To construct unbiased estimating equations for estimating the unknown regression coefficients, for each of these blockwise imputations, we consider their corresponding moment conditions, as follows. For any $r \in [1:R]$, $k \in \mathcal{G}(r)$, and $i \in \mathcal{D}_2$, we consider

$$\boldsymbol{h}_{irk}(\boldsymbol{\beta}) = I(\xi_i = r)\{y_i - (\boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta}\} \cdot \boldsymbol{X}_{ia(k)}^{(k)}, \tag{2.2}$$

where y_i is the response of the ith sample, and $\boldsymbol{X}_{ia(k)}^{(k)}$ is a subvector of $\boldsymbol{X}_i^{(k)}$ consisting of elements corresponding to all covariates observed in Group k. Under the linear regression model, whenever ξ_i is independent of all the covariates (MCAR), or depends only on the observed covariates (MAR), it can be shown that $E\{\boldsymbol{h}_{irk}(\boldsymbol{\beta})\} = \boldsymbol{0}$ (Xue and Qu, 2021). Intuitively, the construction of $\boldsymbol{h}_{irk}(\boldsymbol{\beta})$ is inspired by the score function under the linear regression model, which is still expected to be zero after the blockwise imputations. In addition, note that for different $k_1, k_2 \in \mathcal{G}(r)$ or for different imputations, the dimension of their corresponding equation (2.2) may be different, because the subset $a(k)$ varies with k.

Integrating all missing groups and imputations, we can define a system of unbiased estimating equations as

$$\boldsymbol{g}(\boldsymbol{\beta}) := \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} \begin{bmatrix} \hat{\theta}_1^{-1} \boldsymbol{h}_{i1}(\boldsymbol{\beta}) \\ \vdots \\ \hat{\theta}_R^{-1} \boldsymbol{h}_{iR}(\boldsymbol{\beta}) \end{bmatrix} = 0, \tag{2.3}$$

where $\hat{\theta}_r = |\mathcal{D}_2 \cap \mathcal{S}(r)|/|\mathcal{D}_2|$ is an estimate of the observed rate for the rth group among $\mathcal{D}_2$, and $\boldsymbol{h}_{ir}(\boldsymbol{\beta})$ is a vector combining the components of the vectors in $\{\boldsymbol{h}_{irk}(\boldsymbol{\beta})\}_{k \in \mathcal{G}(r)}$, for $r \in [1 : R]$. In particular, $\boldsymbol{g}(\boldsymbol{\beta})$ is a vector of dimension $M_g = \sum_{r=1}^{R} \sum_{k \in \mathcal{G}(r)} |a(k)|$, which may be larger than p. This overspecification is helpful in terms of making full use of the information contained in all the missing patterns and the available observations. Nonetheless, it is shown in Section S5 of the Supplementary Material (Lemmas 1 and 2) that, under a wide range of settings, the above system of estimating equations leads to a feasible set that contains the true coefficient vector $\boldsymbol{\beta}$ with high probability.

However, the random vectors $\boldsymbol{X}_i^{(k)}$ required by (2.2) and (2.3) are not fully observed. Instead, we use the imputed observations $\widehat{\boldsymbol{X}}_i^{(k)}$ as an approximation. Specifically, we define the imputed counterpart of $\boldsymbol{h}_{irk}(\boldsymbol{\beta})$ as

$$\widehat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta}) = I(\xi_i = r)\{y_i - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta}\} \cdot \widehat{\boldsymbol{X}}_{ia(k)}^{(k)}, \tag{2.4}$$

and define the imputed estimating function as

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} \begin{bmatrix} \hat{\theta}_1^{-1} \widehat{\boldsymbol{h}}_{i1}(\boldsymbol{\beta}) \\ \vdots \\ \hat{\theta}_R^{-1} \widehat{\boldsymbol{h}}_{iR}(\boldsymbol{\beta}) \end{bmatrix}, \tag{2.5}$$

where $n = |\mathcal{D}_2|$, and $\widehat{\boldsymbol{h}}_{ir}(\boldsymbol{\beta})$ is a vector combining the components of the vectors in $\{\widehat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta})\}_{k \in \mathcal{G}(r)}$, for each $r \in [1 : R]$.

Finally, respecting the underlying sparsity of the coefficient vector $\boldsymbol{\beta}$, we define the proposed estimator as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_1, \quad \text{subject to } \|\boldsymbol{g}_n(\boldsymbol{\beta})\|_\infty \leq \lambda, \tag{2.6}$$

where $\lambda > 0$ is a tuning parameter. In Section 3, we obtain the theoretical value for λ up to a constant factor, such that the associated optimizer $\widehat{\boldsymbol{\beta}}$ is a consistent estimator. In practice, we recommend using cross-validation to determine the optimal choice of λ. See Section 5 for more details about the numerical implementation of (2.6).

2.3. Bias-correction based on the projected estimating equations

Although the proposed estimator $\widehat{\boldsymbol{\beta}}$ performs well in terms of point estimation, it is actually biased, and cannot be used directly to develop powerful inference procedures, such as confidence intervals and statistical tests. In

this subsection, we propose a novel projected estimation equation approach incorporating both the unsupervised and the supervised samples, and construct bias-corrected estimators that are asymptotically normally distributed around the true coefficients.

From the imputed estimating function $\boldsymbol{g}_n(\boldsymbol{\beta})$ in (2.5), we define $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ as a subvector of $\boldsymbol{g}_n(\boldsymbol{\beta})$, where we replace each $\widehat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta})$ in (2.4) with its subvector

$$\widehat{\boldsymbol{h}}_{irk}^*(\boldsymbol{\beta}) = I(\xi_i = r)\{y_i - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta}\} \cdot \boldsymbol{X}_{ia(r,k)}^{(k)} = I(\xi_i = r)\{y_i - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta}\} \cdot \boldsymbol{X}_{ia(r,k)}.$$

The dimension of $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ is thus $\sum_{r=1}^R \sum_{k\in\mathcal{G}(r)} |a(r,k)|$. Note that $\widehat{\boldsymbol{h}}_{irk}^*(\boldsymbol{\beta})$ involves only imputed values in $\widehat{\boldsymbol{X}}_i^{(k)}$, whereas the remaining part in $\widehat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta})$ contains imputed values in $\widehat{\boldsymbol{X}}_i^{(k)}$ and $\widehat{\boldsymbol{X}}_{ia(k)\setminus a(r,k)}^{(k)}$, where $\widehat{\boldsymbol{X}}_{ia(k)\setminus a(r,k)}^{(k)}$ is a subvector of $\widehat{\boldsymbol{X}}_i^{(k)}$ consisting of the covariates indexed by $a(k)\setminus a(r,k)$. We have two reasons for using $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ instead of $\boldsymbol{g}_n(\boldsymbol{\beta})$. First, from our theoretical analysis, $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ contributes less error caused by imputation to the final debiased estimator. Second, it significantly simplifies our numerical implementation and improves the finite-sample performance, especially in the optimization (2.9) below.

Based on the initial estimator $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{g}_n^*(\boldsymbol{\beta})$, we propose a bias-corrected estimator $\tilde{\beta}_j$ of β_j for each $j \in [1:p]$, defined as the root of the projected estimating function

$$\hat{S}_j(\widehat{\boldsymbol{\beta}}_j^*) = 0, \tag{2.7}$$

where $\widehat{\boldsymbol{\beta}}_j^* = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{j-1}, \beta_j, \widehat{\beta}_{j+1}, \ldots, \widehat{\beta}_p)^\top$, β_j and $\hat{\beta}_j$ are the jth elements of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, respectively, and

$$\hat{S}_j(\boldsymbol{\beta}) = \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}_n^*(\boldsymbol{\beta}). \tag{2.8}$$

Here, the equation (2.7) is treated as a univariate equation of the scalar β_j, and the projection vector $\hat{\boldsymbol{v}}_j$ is defined as the solution to the following optimization problem:

$$\hat{\boldsymbol{v}}_j = \operatorname*{argmin}_{\boldsymbol{v}} \boldsymbol{v}^\top \boldsymbol{W}_n \boldsymbol{v}, \quad \text{subject to } \|\boldsymbol{v}^\top \boldsymbol{G}_n - \boldsymbol{e}_j\|_\infty \le \lambda', \tag{2.9}$$

where $\lambda' > 0$ is a tuning parameter, $\boldsymbol{e}_j \in \mathbb{R}^p$ has 1 as its jth element and is zero otherwise, $\boldsymbol{W}_n$ is a block-diagonal matrix consisting of the sub-matrices

$$\frac{|\mathcal{D}_2|^2}{|\mathcal{D}_2 \cap \mathcal{S}(r)|^2} \sum_{i\in\mathcal{D}_2} I\{\xi_i = r\} \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^\top,$$

ordered first by $r \in [1:R]$ and then by $k \in \mathcal{G}(r)$, and

$$\boldsymbol{G}_n = \frac{d}{d\boldsymbol{\beta}} \boldsymbol{g}_n^*(\boldsymbol{\beta}) = \frac{1}{|\mathcal{D}_2|} \sum_{i\in\mathcal{D}_2} \begin{bmatrix} \hat{\theta}_1^{-1} d\widehat{\boldsymbol{h}}_{i1}^*(\boldsymbol{\beta})/d\boldsymbol{\beta} \\ \vdots \\ \hat{\theta}_R^{-1} d\widehat{\boldsymbol{h}}_{iR}^*(\boldsymbol{\beta})/d\boldsymbol{\beta} \end{bmatrix}. \tag{2.10}$$

Here in (2.10), for each $r \in [1 : R]$, we have $d\widehat{\boldsymbol{h}}^*_{ir}(\boldsymbol{\beta})/d\boldsymbol{\beta} \in \mathbb{R}^{m'_r \times p}$ with $m'_r = \sum_{k\in\mathcal{G}(r)} |a(r,k)|$ consisting of submatrices $\{I(\xi_i = r)\boldsymbol{X}_{ia(r,k)}(\widehat{\boldsymbol{X}}_i^{(k)})^\top\}_{k\in\mathcal{G}(r)}$ combined by row. Importantly, in (2.10) and (2.9), the unsupervised samples are implicitly used to construct the optimal projection direction $\hat{\boldsymbol{v}}_j$ using the imputed variables. Moreover, in Section 3, we show that having a sufficiently large set of unsupervised samples $\mathcal{D}_1$, and being able to incorporate the information contained in $\mathcal{D}_1$, is necessary to reduce the bias and to obtain the asymptotically normal estimator $\tilde{\beta}_j$.

Remark 1. In Section 3, a theoretical value for the tuning parameter λ' in the quadratic optimization problem (2.9) is obtained, up to a constant factor. For numerical implementations, in Section 5, we propose a practical iterative procedure for determining an appropriate value for λ' that exhibits good numerical performance across various settings.

The rationale behind the projected estimating function in (2.8) is evident from a bias-variance analysis for the estimator $\tilde{\beta}_j$. Specifically, the projected estimating function is carefully constructed using the projection vector $\hat{\boldsymbol{v}}_j$ defined in (2.9), such that the bias term of $\tilde{\beta}_j$ is dominated by a stochastic error, introduced below. Denote $\widetilde{\boldsymbol{\beta}}^*_j = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{j-1}, \tilde{\beta}_j, \widehat{\beta}_{j+1}, \ldots, \widehat{\beta}_p)^\top$. By the Taylor expansion,

$$\begin{aligned} 0 = \hat{S}_j(\widetilde{\boldsymbol{\beta}}^*_j) &= \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}^*_n(\widehat{\boldsymbol{\beta}}^*_j) + \hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j \cdot (\tilde{\beta}_j - \beta_j) \\ &= \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}^*_n(\boldsymbol{\beta}) + \hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}}^*_j - \boldsymbol{\beta}) + \hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j \cdot (\tilde{\beta}_j - \beta_j), \end{aligned} \tag{2.11}$$

which can be rewritten as

$$\tilde{\beta}_j - \beta_j = - \underbrace{\frac{\hat{\boldsymbol{v}}_j^\top \boldsymbol{g}^*_n(\boldsymbol{\beta})}{\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j}}_{\text{Stochastic Error}} - \underbrace{\frac{\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}}^*_j - \boldsymbol{\beta})}{\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j}}_{\text{Remaining Bias}}. \tag{2.12}$$

The estimation error $\tilde{\beta}_j - \beta_j$ is decomposed into two parts. The first term in (2.12) is a stochastic error, which is asymptotically normal with variance determined by $\hat{\boldsymbol{v}}_j^\top \boldsymbol{W}_n \hat{\boldsymbol{v}}_j$, and the remaining bias is bounded by

$$\left| \frac{\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}}^*_j - \boldsymbol{\beta})}{\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j} \right| \le \frac{\|(\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n)_{-j}\|_\infty \|(\widehat{\boldsymbol{\beta}}^*_j - \boldsymbol{\beta})_{-j}\|_1}{1 - |\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n \boldsymbol{e}_j - 1|} \le \frac{\|\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1}{1 - \|\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty}, \tag{2.13}$$

using Hölder's inequality. As a result, the remaining bias is dominated by the stochastic error, because the factor $\|\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty$ is well controlled by (2.9), and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ is sufficiently small.

From the above argument, the constrained optimization problem (2.9) is rooted in a bias–variance trade-off: It aims to find a projection vector $\hat{\boldsymbol{v}}_j$ that controls $\|\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty$ in (2.13) to ensure the remaining bias in (2.12) is

negligible with respect to the stochastic error, while reducing the variance of the stochastic error, by minimizing $\hat{\boldsymbol{v}}_j^\top \boldsymbol{W}_n \hat{\boldsymbol{v}}_j$, to obtain a more efficient estimator.

Remark 2. For general missing data problems, there are likelihood-based approaches where missing values are marginalized under distributional assumptions (Garcia, Ibrahim and Zhu, 2010; Ibrahim, Lipsitz and Chen, 1999; Chen, Prentice and Wang, 2014). In particular, the expectation–maximization (EM)-based estimating equation method also constructs estimating functions based on missing data (Elashoff and Ryan, 2004). However, the proposed projected estimating equations and the EM-based estimating equations are conceptually different. First, the proposed method does not need to specify distributions for all the variables. Second, the projected estimating equations are carefully designed to correct the bias of our initial estimator. In contrast, the EM-based estimating equations are the derivatives of the log-likelihood function with respect to the parameters (Elashoff and Ryan, 2004).

Remark 3. In our construction of $\tilde{\beta}_j$, we mainly correct for the bias due to $\widehat{\boldsymbol{\beta}}_{-j}$, as in (2.13), rather than the bias due to $\{\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} : j \in a(r)^c, k \in \mathcal{G}(r), 1 \le r \le R\}$ from the imputation procedure. In general, the bias of the final estimator $\widetilde{\beta}_j$ stems partially from the estimation error of the conditional expectation $E(X_{ij}|\boldsymbol{X}_{ia(r,k)})$, defined as the difference between $\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$ and $\boldsymbol{\gamma}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$ under the linear assumption $E(X_{ij} \mid \boldsymbol{X}_{ia(r,k)}) = \boldsymbol{\gamma}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$. The estimation error can be well controlled by $\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2$, which contains both the bias and the variance of $\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}$. To obtain a small estimation error, we leverage both unsupervised and supervised samples to ensure that $\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2$ is small with high probability.

3. Theoretical Justifications

This section provides theoretical justifications for the proposed inference procedures by studying the properties of the proposed estimator $\widehat{\boldsymbol{\beta}}$ and its bias-corrected counterpart $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top$. For technical reasons, we assume for simplicity that the blockwise imputation step (2.1) is performed using the unsupervised samples $\mathcal{D}_1$ and a fixed portion of the supervised samples $\mathcal{D}_2$ that preserve the blockwise missing pattern (i.e., the number of groups and the missing variables in each group). On the other hand, the construction of the estimators $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ is based on the imputed observations of the other portion of the supervised samples $\mathcal{D}_2$. In practice, however, splitting the supervised samples $\mathcal{D}_2$ into two parts is not needed, and the proposed method works well numerically when all the samples are used for imputation and inference; see the numerical results in Sections 5 and 6.

We first introduce the notation and assumptions for the theoretical results. For any $r \in [1 : R]$, $k \in \mathcal{G}(r)$ and $i \in \mathcal{D}$, we define $\boldsymbol{\Sigma}^{(r,k)} = E\{I(\xi_i = r)\boldsymbol{X}_i^{(k)}(\boldsymbol{X}_i^{(k)})^\top\} \in \mathbb{R}^{p\times p}$. Recall that $\xi_i \in [1 : R]$ denotes the random group label

of the ith sample, $a(k)^c$ is the index set of the missing covariates in Group k, $a(k)$ is the index set of the observed covariates in Group k, and $a(r,k)$ is the index set of the covariates observed in both Groups r and k. We also denote $N_r = |\mathcal{D}_1 \cap \mathcal{S}(r)|$ as the number of unsupervised samples in Group r, $n_r = |\mathcal{D}_2 \cap \mathcal{S}(r)|$ as the number of supervised samples in Group r, and $N = |\mathcal{D}_1|$.

For the missingness mechanism, we assume the following:

(A1) The random group label ξ_i is independent of all covariates or depends only on covariates observed in all groups, and the response is MCAR.

(A2) For the missing patterns, we assume R is a finite integer, and for all $r \in [1 : R]$ and $k \in \mathcal{G}(r)$, we have $|a(r)|/p, a(r,k)/p \in [C_1, C_2]$ and $n_r/n, N_r/N \in [c_1, c_2]$, with probability at least $1 - p^{-c}$ for some constants $0 < C_1 < C_2 < 1$, $0 < c_1 < c_2 < 1$, and $c > 0$.

The assumption for the random group label in (A1) implies the missingness mechanisms of the covariates are either MCAR or MAR, because the missingness (or group assignments) is completely random or can be fully explained by completely observed variables.

Assumption (A2) is mild, because it essentially ensures that the missing patterns are finite and balanced. For the design covariates, and the regression coefficient vector $\boldsymbol{\beta}$, we assume the following:

(A3) Each $\boldsymbol{X}_i$, for $i \in \mathcal{D}$, is an independent centered sub-Gaussian random vector with $\boldsymbol{\Sigma} = E(\boldsymbol{X}_i \boldsymbol{X}_i^\top)$ satisfying $C^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le C$, for some absolute constant $C > 1$, and $\boldsymbol{\gamma}_j = \operatorname{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} E(X_{ij} - \boldsymbol{\gamma}^\top \boldsymbol{X}_{i,-j})^2$ satisfies $\|\boldsymbol{\gamma}_j\|_0 \le s$, for each $j \in [1 : p]$;

(A4) $\boldsymbol{\beta}$ satisfies $\|\boldsymbol{\beta}\|_2 \le C$, for some absolute constant $C > 0$.

(A5) There exists some $r \in [1 : R]$, $k_1, k_2 \in \mathcal{G}(r)$ and some constant $c_0 > 0$, such that $\lambda_{\min}(\boldsymbol{\Sigma}^{(r,k)}_{a(k),a(k)}) \ge 7c_0 > c_0 \ge \lambda_{\max}(\boldsymbol{\Sigma}^{(r,k)}_{a(k),a(k)^c})$, for $k = k_1, k_2$, and $a(k_1) \cup a(k_2) = [1 : p]$. In Assumption (A3), the sub-Gaussian condition includes many important cases, such as Gaussian, bounded, and binary covariates, or any combinations of them. This makes our proposed method applicable to many practical settings. The sparsity condition on the best linear predictor coefficient γ_j ensures the quality of the Lasso-based imputation step, which essentially requires a sparse conditional dependence structure among the covariates. For example, when $\boldsymbol{X}_i \overset{i.i.d.}{\sim} N(0, \boldsymbol{\Sigma})$, this condition is equivalent to a sparse Gaussian graph condition, requiring each row of $\boldsymbol{\Sigma}^{-1} = (\omega_{ij})$ to be s-sparse.

Assumption (A5) requires the existence of two groups $\{k_1, k_2\} \subseteq \mathcal{G}(r)$, such that each covariate is observed in one of these two groups. However, the eigenvalue

condition $\lambda_{\min}(\mathbf{\Sigma}^{(r,k)}_{a(k),a(k)}) \geq 7c_0 > c_0 \geq \lambda_{\max}(\mathbf{\Sigma}^{(r,k)}_{a(k),a(k)^c})$ requires the existence of a pair of groups $(r,k) \in [1:R] \times \mathcal{G}(r)$ such that, for each $i \in \mathcal{S}(r)$, the subvector $\boldsymbol{X}^{(k)}_{ia(k)}$ of the imputed vector $\boldsymbol{X}^{(k)}_i$ does not contain variables that are highly correlated within themselves, or with the variables in $\boldsymbol{X}^{(k)}_{ia(k)^c}$. This condition essentially ensures that each covariate is sufficiently informative. In Section S8 of the Supplementary Material, a more interpretable sufficient condition is obtained under the Gaussian design.

The following theorem concerns the convergence rates of the estimator $\widehat{\boldsymbol{\beta}}$ in (2.6).

Theorem 1. *Suppose* (A1) *to* (A5) *hold,* $\log p \ll \min\{N, n\}$, *and* $s \ll \min\{\sqrt{n/\log p}, (n+N)/\log p\}$. *For sufficiently large* (n,p), *if we choose* $\tau \asymp \sqrt{\log p/(n+N)}$ *in* (2.1) *and* $\lambda \asymp \sqrt{\log p/n} + s\sqrt{\log p/(n+N)}$ *in* (2.6), *then* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \lesssim s\lambda$ *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim s^{1/2}\lambda$ *hold with probability at least* $1 - p^{-c}$, *for some absolute constant* $c > 0$.

Some remarks about Theorem 1 are in order. First, our theorem shows that the rate of convergence under the ℓ_2-norm is bounded by $\sqrt{s\log p/n} + s^{3/2}\sqrt{\log p/(n+N)}$. The first term $\sqrt{s\log p/n}$ is the ordinary estimation error for the Lasso or Dantzig selector type of estimators, whereas the second term $s^{3/2}\sqrt{\log p/(n+N)}$ comes from the estimation error of the conditional expectation in the BI step for the missing covariates. Intuitively, the estimation error of the conditional expectation depends on both N and n, because the BI step uses both the supervised and the unsupervised samples. In contrast, the estimation error of the Lasso or Dantzig selector depends only on n, because only the imputed supervised samples are used in the estimating equations (2.5).

Second, compared with the minimax optimal rate $\sqrt{s\log p/n}$ for estimating $\boldsymbol{\beta}$ with complete observations of n samples with $\lambda \asymp \sqrt{\log p/n}$ (Verzelen, 2012), the above error rate has an additional term $s^{3/2}\sqrt{\log p/(n+N)}$ under $\lambda \asymp \sqrt{\log p/n} + s\sqrt{\log p/(n+N)}$. This extra error term and the different choice of tuning parameters reflect the cost of imputing the missing variables; see also Chandrasekher, Alaoui and Montanari (2020) for similar phenomena in the imputation of unstructured missing data using the Lasso method. However, Theorem 1 also implies that when the number of unsupervised samples is sufficiently large, that is, when $N \gtrsim s^2 n$, the estimation error of the conditional expectation is dominated by the estimation error $\sqrt{s\log p/n}$, and the estimator $\widehat{\boldsymbol{\beta}}$ achieves the minimax optimal rate for complete observations of n samples. In other words, our method benefits from the extra unsupervised samples to improve the estimation. Nevertheless, note that even in the presence of a far greater number of unsupervised samples ($N \gg n$), the convergence rate cannot be better than $\sqrt{s\log p/n}$. After all, there are only n observations of the response variable, rather than n complete samples.

Third, unlike many existing inferential methods for missing data, such as those of Cai, Cai and Zhang (2016), Kundu, Tang and Chatterjee (2019), and Yu et al. (2020), our method does not require fully observed samples. In other words, each sample in the data set may have missing variables, which precludes using existing methods for fully observed data. In contrast, our method should work, as long as $|\mathcal{G}(r)| \geq 1$ and the missing groups are finite and asymptotically balanced.

The proof of Theorem 1 is involved and quite different to existing works that analyze the risk bound of the Dantzig selector or the Lasso estimator for the linear regression model with complete data (Candes and Tao, 2007; Bickel, Ritov and Tsybakov, 2009). The detailed proof can be found in Section S5 of the Supplementary Material. In particular, as a key component of our theoretical analysis, we develop a novel restricted singular value inequality that accounts for the blockwise-imputed samples, and plays a similar role to the restricted eigenvalue condition (Raskutti, Wainwright and Yu, 2010) or the restricted strong convexity property (Negahban et al., 2012; Negahban and Wainwright, 2012) needed to analyze of high-dimensional ℓ_1-penalized estimators. This inequality, proved in Section S7.4 of the Supplementary Material, could be of independent interest.

Proposition 1. *Under the conditions of Theorem* 1*, there exist some* $r \in [1:R]$ *and* $k \in \mathcal{G}(r)$*, such that, with probability at least* $1-p^{-c}$ *for some absolute constant* $c > 0$*,*

$$\inf_{\substack{\|\boldsymbol{u}\|_2=1, \boldsymbol{u}\in E_s(p) \\ \|\boldsymbol{u}_{a(k)}\|_2 \geq 1/2}} \left| n_r^{-1} \sum_{i=1}^{n} I\{\xi_i = r\} \left(\frac{\boldsymbol{u}_{a(k)}}{\|\boldsymbol{u}_{a(k)}\|_2} \right)^\top \widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{u} \right| \geq c_0, \tag{3.1}$$

for some constant $c_0 > 0$*, where* $E_s(p) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_2 = 1, \|\boldsymbol{\delta}_{S^c}\|_1 \leq \|\boldsymbol{\delta}_S\|_1$*, for some set* $S \subset [1:p]$ *with* $|S| \leq s\}$*, and* S^c *represents the complement of set* S*.*

Our next theorem establishes the asymptotic normality of the bias-corrected estimator $\tilde{\beta}_j$, supporting the asymptotic validity of the confidence intervals and the statistical tests proposed in Section 4. We need the following condition, ensuring the existence of a true projection vector satisfying the constraint in (2.9), with high probability:

(A6) For $\boldsymbol{G} = d\boldsymbol{g}^*(\boldsymbol{\beta})/d\boldsymbol{\beta}$ with $\boldsymbol{g}^*(\boldsymbol{\beta})$ being the population counterpart of $\boldsymbol{g}_n^*(\boldsymbol{\beta})$, we have $\lambda_{\min}(E\{\boldsymbol{G}\}) \geq c$, for some absolute constant $c > 0$,

Theorem 2. *Suppose the conditions of Theorem* 1 *and* (A6) *hold, and* $N \gtrsim n \log p$*. If we choose* $\lambda' \asymp \sqrt{\log p/n}$ *and* $s \ll \min\{\sqrt{n}/\log p, \sqrt{N/n \log p}\}$*, then, for each* $j \in [1:p]$*, we have*

$$\frac{n(\tilde{\beta}_j - \beta_j)}{s_j} = AB + D, \tag{3.2}$$

where s_j is defined in (4.1)*, $A \to 1$ and $D \to 0$ in probability, and $B|\widehat{X} \to N(0,1)$ in distribution, in which $\widehat{X} = \{\widehat{\boldsymbol{X}}_i^{(k)}\}_{i\in\mathcal{D}_2}$ is the set of all imputed observations.*

Theorem 2 shows that to obtain an asymptotically normally distributed estimator, we needa sufficiently large set of unsupervised samples for both the blockwise imputation and the bias correction. Specifically, from our proof of Theorem 2 (such as Lemma 6 in the Supplementary Material), it seems that, under the current analytical framework, the condition $N \gtrsim n \log p$ is likely necessary for constructing nearly unbiased estimators with efficiency competitive to $\tilde{\beta}_j$. In addition, the condition $s \ll \sqrt{N/n \log p}$ ensures that the imputation error is $o(n^{-1/2})$, whereas the more standard condition $s \ll \sqrt{n}/\log p$ implies that the remaining bias in (2.12) after the bias-correction step is negligible.

These conditions are explained as follows. On the one hand, additional unsupervised samples are needed to achieve desirable imputation quality, that is, to ensure the imputation error is dominated by the estimation error for $\widehat{\boldsymbol{\beta}}$. Intuitively, if the imputation error dominates the estimation error in the bias of $\widehat{\boldsymbol{\beta}}$, then such a bias is intrinsic, and may not be removed by any approach based on the imputed data. On the other hand, the unsupervised samples help to reduce bias: the proposed projected estimating equation approach incorporates both the unsuperivsed and the supervised samples to jointly determine the best projection direction in (2.9) for bias correction. We also provide theoretical results when we have only supervised samples in Section S4 of the Supplementary Material, showing that the convergence rate of the proposed estimator is faster for both supervised and unsupervised samples than it is for only supervised samples.

4. Confidence Intervals and Statistical Tests

In this section, we develop asymptotically valid confidence intervals and statistical tests for each coefficient β_j, with $j \in [1:p]$. As shown in Section 3, by carefully analyzing the bias-corrected estimator $\tilde{\beta}_j$, conditional on the imputed covariates, under mild regularity conditions, $\tilde{\beta}_j$ is asymptotically normally distributed with variance $s_j^2/|\mathcal{D}_2|^2$, where

$$s_j^2 = \sum_{i\in\mathcal{D}_2} \sum_{k\in\mathcal{G}(r),1\le r\le R} \frac{|\mathcal{D}_2|^2\sigma_{r,k}^2}{|\mathcal{D}_2 \cap \mathcal{S}(r)|^2} I(\xi_i = r)(\hat{\boldsymbol{v}}_{j,rk}^\top \boldsymbol{X}_{ia(r,k)})^2, \tag{4.1}$$

$\sigma_{r,k}^2 = \sigma^2 + \boldsymbol{\beta}_{a(r)^c}^\top E\{\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}(\boldsymbol{\epsilon}_{ia(r)^c}^{(k)})^\top\}\boldsymbol{\beta}_{a(r)^c}$, $\boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \in \mathbb{R}^{|a(r)^c|}$ is the residual term of the ith sample in the regression model of $\boldsymbol{X}_{ia(r)^c}$, with $\boldsymbol{X}_{ia(r,k)}$ as covariates, and $\hat{\boldsymbol{v}}_{j,rk} \in \mathbb{R}^{|a(r,k)|}$, with $r \in [1:R]$ and $k \in \mathcal{G}(r)$, is the subvector of the projection vector $\hat{\boldsymbol{v}}_j$ corresponding to the estimating functions in $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ associated with

Group $k \in \mathcal{G}(r)$. Consequently, for any given $j \in [1:p]$, an asymptotically $(1-\alpha)$-level confidence interval for β_j can be constructed as $\mathrm{CI}_\alpha(\beta_j) = [\tilde{\beta}_j - z_{\alpha/2}\widehat{s}_j/|\mathcal{D}_2|, \tilde{\beta}_j + z_{\alpha/2}\widehat{s}_j/|\mathcal{D}_2|]$, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the upper $\alpha/2$-quantile of the standard normal distribution,

$$\widehat{s}_j^2 = \widehat{\sigma}^2 \sum_{i \in \mathcal{D}_2} \sum_{k \in \mathcal{G}(r), 1 \le r \le R} \frac{|\mathcal{D}_2|^2}{|\mathcal{D}_2 \cap \mathcal{S}(r)|^2} I(\xi_i = r)(\hat{\boldsymbol{v}}_{j,rk}^\top \boldsymbol{X}_{ia(r,k)})^2, \tag{4.2}$$

and $\widehat{\sigma}^2$ is some reasonable estimator for $\max_{k,r} \sigma_{r,k}^2$ (see Section S2 of the Supplementary Material).

Along with the above confidence interval, we also construct an asymptotically valid statistical test for the null hypothesis $H_0 : \beta_j = b_j$, for any $b_j \in \mathbb{R}$. Specifically, we define a test statistic $T_j = |\mathcal{D}_2|(\tilde{\beta}_j - b_j)/\widehat{s}_j$. Then, an asymptotically α-level two-sided test rejects H_0 whenever $|T_j| > z_{\alpha/2}$. With these component-wise test statistics, one can also construct tests for the global null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the multiple simultaneous hypotheses $H_{0j} : \beta_j = 0, j \in [1:p]$. For example, to test the global null hypothesis, we could adopt the maximum-type test statistic $M = \max_{1 \le j \le p} T_j^2$, and compare its empirical values with the quantile of the Gumbel distribution given in Theorem 1 of Ma, Cai and Li (2021).

To test simultaneous null hypotheses while controlling for false discovery rates, we can apply the modified Benjamini–Hochberg procedure in Javanmard and Javadi (2019) and Ma, Cai and Li (2021) to design covariates that are weakly correlated, or the Benjamini–Yekutieli procedure (Benjamini and Yekutieli, 2001) if the design covariates are arbitrarily correlated. The theoretical validity of these simultaneous inference procedures follows from the arguments in Javanmard and Javadi (2019) and Ma, Cai and Li (2021).

5. Simulation

We provide simulation studies to compare the proposed method with existing methods, including the debiased Lasso method (Javanmard and Montanari, 2014) with complete cases, the Lasso projection method (van de Geer et al., 2014; Zhang and Zhang, 2014) with complete cases, the debiased Lasso method with single regression imputation, and the Lasso projection method with the single regression imputation. Here, "single regression imputation" refers to predicting missing values using linear regressions, with observed variables as predictors (Baraldi and Enders, 2010; Zhang, 2016; Campos et al., 2015).

To implement of the proposed method, we use the R packages `glmnet`[1], `Rglpk`[2], and `osqp`[3] to solve the minimization problem in (2.1), the linear pro-

[1] `https://cran.r-Bproject.org/web/packages/glmnet/index.html`
[2] `https://cran.r-Bproject.org/web/packages/Rglpk/index.html`
[3] `https://cran.r-Bproject.org/web/packages/osqp/index.html`

gramming problem in (2.6), and the quadratic programming problem in (2.9), respectively. The parameters τ and λ are determined by cross-validation, which might not achieve the desired theoretical convergence rates. This is one limitation of the proposed method. We let $\lambda' = 0.1(\log p/n)^{1/2}$, and scale it up if there is no solution to the quadratic programming problem in (2.9). The R functions of the proposed method have been made publicly available online at `https://github.com/feixue-stat/Inference_blockmissing`. We use the R code in `https://web.stanford.edu/~montanar/sslasso/` to implement the debiased Lasso method. For the Lasso projection method, we apply the R package `hdi`[4].

For each $i \in [1 : (n+N)]$, we simulate $\boldsymbol{X}_i$ independently from a multivariate Gaussian distribution with mean zero and a covariance matrix $\boldsymbol{\Sigma}$, and generate $y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. The relevant covariates share the same signal strength β_s; that is, the nonzero elements in $\boldsymbol{\beta}$ are all equal to β_s. In the following three settings, all samples are randomly assigned to four missing groups. In Settings 1 and 2, we assume MNAR for the covariates from three sources, and the four missing pattern groups are shown in Figure 2. In contrast, we assume MAR in Setting 3, and add one additional data source, where the variables are all observed for each subject. For the missingness of the response, in each setting, the response is MCAR, where only n/N of all samples in each group are observed. This satisfies Assumption (A1).

In each setting, we construct confidence intervals for a relevant covariate with confidence level 95%, and evaluate each method using the coverage rate and average length of the confidence intervals based on 250 replications. Let p_l denote the number of total covariates in the lth data source, and s_l denote the number of relevant covariates in the lth data source, for $l \in [1 : S]$. Recall that s denotes the number of all relevant covariates, specifying the sparsity of the coefficient vector $\boldsymbol{\beta}$. That is, we have s nonzero elements in $\boldsymbol{\beta}$. In addition, recall that n_r denotes the number of supervised samples in the rth missing group, for $r \in [1 : R]$. Then, we have $\sum_{r=1}^{R} n_r = n$.

Setting 1. Let $n = 150$, $p = 200$, $s = 9$, $R = 4$, $S = 3$, $N = 300$, $\beta_s = 0.2$, $n_1 = 30$, $n_2 = 70$, $n_3 = n_4 = 25$, $p_1 = 115$, $p_2 = 45$, $p_3 = 40$, $s_1 = 5$, $s_2 = s_3 = 2$, and $\boldsymbol{\Sigma} = \mathrm{diag}\{\boldsymbol{I}_{p_1}, \boldsymbol{A}\}$, where $\boldsymbol{I}_{p_1}$ is an identity matrix of size p_1, and $\boldsymbol{A}$ is a $(p_2 + p_3) \times (p_2 + p_3)$ exchangeable matrix with diagonal elements one and off-diagonal elements ρ. We let $\rho = 0.1$ or 0.3, and let the covariates be MNAR. Specifically, samples are sequentially randomly assigned into the complete case group with probabilities proportional to $\exp(-10y_i)$ for $1 \le i \le n+N$. Otherwise, they are uniformly assigned to the other three missing groups.

Setting 2. The same as Setting 1, except that $p = 700$ and $p_1 = 615$.

[4]`https://cran.r-Bproject.org/web/packages/hdi/index.html`

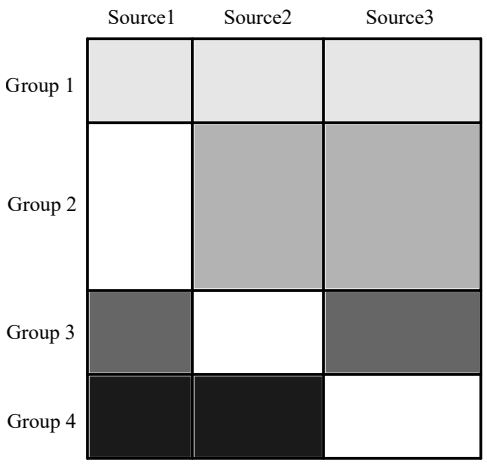

Figure 2. Blockwise missing structure used for simulation.

Setting 3. The same as Setting 1, except that $n = 120$, $S = 4$, $N = 600$, $n_1 = 15$, $n_2 = n_3 = n_4 = 35$, $p_2 = 40$, $p_4 = 5$, $s_1 = 4$, $s_4 = 1$, and $\mathbf{\Sigma} = \text{diag}\{\boldsymbol{I}_{p_1}, \boldsymbol{A}, \boldsymbol{I}_{p_4}\}$. We let the covariates be MAR. Specifically, samples are sequentially randomly assigned into the complete case group with probabilities proportional to $\exp(-10d_i)$ for $1 \leq i \leq n + N$, where d_i is the sum of the ith samples of covariates in the fourth source of data. Otherwise, they are uniformly assigned to the other three missing groups. The missing patterns of the covariates in Sources 1–3 are the same as that in Figure 2, and covariates in Source 4 are all observed.

The results of Settings 1–3 are provided in Table 1, where ρ represents the correlations between covariates. We use different ρ to investigate the performance under various strengths of dependence between the covariates. In Table 1, the proposed method outperforms existing methods across all settings in terms of coverage rate. In Setting 1, 80% of samples have missing covariates, and the missingness is MNAR. Even so, as shown in Table 1, the coverage of the proposed method is at least 42.0% and 19.7% more than that of other methods when $\rho = 0.1$ and $\rho = 0.3$, respectively.

In Setting 2, we consider simulations with potential predictors to mimic the ADNI data in Section 6. The proposed method still produces the largest coverage rate. Moreover, when $\rho = 0.1$, the coverage rate of the proposed method is 94.4%, which is close to 95%. Note that the MNAR missingness mechanism of the covariates in both Settings 1 and 2 violates the MAR assumption (A1), which may explain why the coverage of the proposed method does not achieve 95%. However, there might be other reasons for the lower coverage, such as the limited sample size, missing proportion of responses, and structure of the covariance matrix $\mathbf{\Sigma}$.

Table 1. Simulation results of Settings 1–3. DL-CC: the debiased Lasso method with complete cases. LP-CC: the Lasso projection method with complete cases. DL-SI: the debiased Lasso method with single regression imputation. LP-SI: the Lasso projection method with single regression imputation.

	$\rho = 0.1$		$\rho = 0.3$	
Method	Coverage rate	Average length	Coverage rate	Average length
Setting 1				
Proposed	**0.920**	0.581	**0.876**	0.560
DL-CC	0.264	0.274	0.248	0.291
LP-CC	0.636	0.423	0.644	0.429
DL-SI	0.036	0.140	0.036	0.135
LP-SI	0.648	0.326	0.732	0.362
Setting 2				
Proposed	**0.944**	0.931	**0.908**	0.881
DL-CC	0.000	0.004	0.000	0.008
LP-CC	0.628	0.428	0.668	0.443
DL-SI	0.036	0.146	0.016	0.140
LP-SI	0.804	0.375	0.800	0.380
Setting 3				
Proposed	**0.956**	0.722	**0.956**	0.699
DL-CC	0.260	0.217	0.308	0.229
LP-CC	0.964	1.229	0.924	1.228
DL-SI	0.116	0.191	0.116	0.173
LP-SI	0.356	0.227	0.404	0.252

Setting 3 focuses on MAR and contains additional unsupervised samples. In Table 1, the proposed method and the Lasso projection method with complete cases (LP-CC) both achieve desirable coverage. However, the average length of the confidence intervals of the proposed method is much smaller than that of the LP-CC, indicating that the confidence intervals of the proposed method are more accurate.

In Table 4 of the Supplementary Material, we compare the empirical bias and the empirical standard deviation of each method under Setting 3, and with the multivariate imputation by chained equations (MICE) method. The results show that the proposed estimator has a much smaller empirical standard deviation than that of the LP-CC, and that MICE-based methods produce much larger biases than that of the proposed method. Moreover, although in Table 1 the confidence intervals of LP-SI have poor coverage, Table 4 of the Supplementary Material shows that its point estimator has the smallest mean squared error (squared bias plus variance). In addition, we provide absolute values of empirical biases of $\widehat{\boldsymbol{\beta}}_j$ and $\widetilde{\boldsymbol{\beta}}_j$, and histograms of $\widehat{\boldsymbol{\beta}}_j$ for the jth covariate under Setting 3 in the Supplementary Material, showing that the empirical bias of $\widehat{\boldsymbol{\beta}}_j$ is much larger than that of $\widetilde{\boldsymbol{\beta}}_j$, and that the empirical distribution of $\widehat{\boldsymbol{\beta}}_j$ is right-skewed.

For the effects of the degree of correlations (ρ) between the covariates on the proposed method, Table 1 shows that the coverage rate of the proposed method is lower for larger ρ under Settings 1 and 2, and Table 4 of the Supplementary Material shows that the proposed method has slightly greater bias for larger ρ under Setting 3.

6. Real-Data Application

In this section, we apply the proposed method to the ADNI data set, which contains multisource measurements: MRI, PET imaging, gene expressions, and cognitive tests (Mueller et al., 2005). Of the latter tests, the mini-mental state examination is often used to diagnose Alzheimer's Disease (AD) (Chapman et al., 2016). It is therefore important to identify the imaging and gene expression features that are associated with and can predict the score of the mini-mental state examination. To identify biomarkers associated with AD, we use the score of the mini-mental state examination as our response variable, and treat the MRI, PET, and gene expression variables as predictors.

Specifically, the MRI variables contain volumes, surface areas, average cortical thickness, and the standard deviation of the cortical thickness of regions of interest in the brain, which are extracted from the MRIs by the Center for Imaging of Neurodegenerative Diseases at the University of California, San Francisco. To mitigate bias due to different head sizes, we normalize the MRI variables by dividing the region volumes, surface areas, and cortical thicknesses by the whole-brain volume, the total surface area, and the mean cortical thickness of each subject, respectively (Zhou et al., 2014; Kang et al., 2019). The PET variables are standard uptake value ratios of brain regions of interest, that represent metabolic activity, and are provided by the Jagust Lab at the University of California, Berkeley. Gene expression levels at different probes are contributed by Bristol-Myers Squibb laboratories from blood samples of ADNI participants.

Although the ADNI is a longitudinal study, we focus on data collected in the second phase of the study (ADNI-2), at month 48. In total, there are 212 samples, 267 MRI variables, 113 PET variables, and 49,386 gene expression variables. The blockwise missingness emerges when we combine the MRI, PET, and gene expression data. The missing pattern structure is the same as that in Figure 2, with four groups and 69 complete observations. Because of the relatively small sample sizes, we first screen the gene expression variables using marginal correlations based on sure independence screening (Fan and Lv, 2008) retaining 300 gene expression variables. We compute the marginal correlation between the response variable and each gene expression variable based on all available pairs of observations of the two variables.

We first apply the proposed method to all $n = 212$ samples in order to identify the biomarkers associated with the score of the mini-mental state

examination. We test the simultaneous hypotheses $H_{0j} : \beta_j = 0, 1 \le j \le p = 680$, while controlling the false discovery rate (FDR), using the modified Benjamini–Hochberg procedure of Ma, Cai and Li (2021) with the proposed estimators $\tilde{\beta}_j$ and their variance estimators $\widehat{s}_j^2$. The multiple testing procedure assumes that the true alternatives are sparse, and is shown to control the FDR in probability under mild conditions as $n \to \infty$. See Section S9 of the Supplementary Material for more details about the testing procedure.

The biomarkers identified by the methods at the significance level $\alpha = 0.01$ are provided in Table 6 of the Supplementary Material. For the gene expression probes, we provide the corresponding gene names in the table. The proposed method identifies 36 biomarkers, including 19, 2, and 15 variables from the MRI, PET, and gene expressions, respectively. Some of these biomarkers are also selected by other methods. We provide the biomarkers identified by both the proposed method and one of other methods in Table 7 in the Supplementary Material. Although the debiased Lasso using complete cases or using single regression imputation seems to identify many more markers, based on our simulation results, many of the identified markers may be false positives, because the corresponding confidence intervals do not provide the correct coverage probabilities.

Among the associated genes, SFRP1 is selected by all the methods, and is crucial in AD pathogenesis (Esteve et al., 2019). PJA2 is identified only by the proposed method, and has reduced expressions in AD patients than on normal controls. PJA2 has been shown to regulate AD marker genes in mouse hippocampal neuronal cells, indicating its the potential relevance to the pathophysiology of AD (Gong et al., 2020). Among the MRI related markers, "`ST30SV`" is identified by our method as well as DL-SI and LP-SI, and represents the volume of the left inferior lateral ventricle, which is related to AD (Bartos et al., 2019; Ledig et al., 2018). However, only the proposed method identifies "`ST101SV`" and "`ST35TA`", representing the volume of the right pallidum and the average cortical thickness of the left lateral occipital, respectively. Both are shown to be associated with AD (Kautzky et al., 2018; Yang et al., 2019). Finally, the PET biomarker "`CTX_RH_TEMPORALPOLE`," the standardized uptake value of the right temporal pole, is identified only by our method. This agrees with the observation that hypometabolism in the temporal lobe often appears in AD patients (Sanabria-Diaz, Martínez-Montes and Melie-Garcia, 2013).

To show that the multiple sources in the ADNI study contain complementary information, we compare the proposed method with the Lasso using only the MRI, PET, or gene expression variables for prediction. We also compare the proposed method with the naive mean prediction method and the Lasso using only complete observations. The naive mean prediction method uses the sample mean of the response variable, calculated based on training sets for prediction. Specifically, we randomly hide 10% of all values of the response variable as testing

responses 150 times, and apply all the methods to the remaining data. In each replication, we calculate the prediction mean squared error $\sum_{1\le i\le T}(\hat{y}_i - y_i)^2/T$, where y_i is a testing response, $\hat{y}_i$ is the corresponding predicted value, and T is the number of testing responses. We also compute the improvement rates of the proposed method relative to other methods in terms of the prediction mean squared error, which is defined as $(PE_{\mathcal{M}} - PE_{\mathcal{P}})/PE_{\mathcal{P}}$, where $PE_{\mathcal{P}}$ and $PE_{\mathcal{M}}$ denote the averages of the prediction mean squared errors of the proposed method and the method $\mathcal{M}$, respectively, based on the 150 replications.

As shown in Table 2, the proposed estimator $\widehat{\boldsymbol{\beta}}$ produces smaller prediction mean squared errors than other estimators, indicating that the proposed method can achieve higher prediction accuracy than when using data from only one source or when using only complete cases. This implies that using all the data sources (MRI, PET, and Gene) with the proposed method can improve the prediction compared with using a subset of predictors. Note that this is not over-fitting, because the prediction errors in Table 2 are testing errors, rather than training errors. Thus, different data sources in the ADNI study contain complementary information, and the proposed integration method is suitable in that respect.

Specifically, the proposed method reduces the prediction mean squared errors of other methods by at least 10.6%. In particular, the improvement rate with respect to the Lasso method using only gene expression variables or using only complete cases is over 30%. Moreover, the standard deviation of the prediction mean squared errors of the proposed method is smaller than that of other the methods, indicating that the proposed method is more stable. Furthermore, we provide the absolute mean (absolute value of the mean) and standard deviation of $\hat{y}_i - y_i$, for $i = 1, \ldots, T$, in Table 8 of the Supplementary Material. We also provide the squared bias $\sum_{i=1}^{n} I(y_i \in \mathcal{T}) \cdot (\sum_{j=1}^{t_i} \hat{y}_{ij}/t_i - y_i)^2/|\mathcal{T}|$ and variance $\sum_{i=1}^{n} I(y_i \in \mathcal{T}) \cdot \sum_{j=1}^{t_i} (\hat{y}_{ij} - \sum_{j=1}^{t_i} \hat{y}_{ij}/t_i)^2/(t_i|\mathcal{T}|)$ in Table 9 of the Supplementary Material, where n is the total number of samples in the real data, $\mathcal{T}$ is the set of responses that are included in at least one test set, $\hat{y}_{ij}$ is the jth predicted value by a method for y_i in all test sets, and t_i is the total number of the predicted values $\hat{y}_{ij}$ in all test sets. The results show that the proposed method produces the smallest squared bias among all the methods.

In summary, the proposed estimator produces smaller prediction mean squared errors and smaller squared bias than using only one source data or using only complete observations, implying that integrating data from multiple sources and using incomplete observations are critical. Additionally, the proposed method identifies meaningful and important biomarkers not selected by other methods, indicating that the proposed method is more powerful in terms of integrating multimodal data.

Table 2. Averages of prediction mean squared errors based on 150 replications. Proposed ($\widehat{\boldsymbol{\beta}}$): the proposed method with the estimator $\widehat{\boldsymbol{\beta}}$. MRI Lasso, PET Lasso, and Gene Lasso: Lasso method using only MRI, PET, and gene expression variables, respectively. CC Lasso: the Lasso method using only complete cases. Naive mean: using the sample mean of the response variable in the training sets for prediction. SD: standard deviation of prediction mean squared errors calculated based on 150 replications.

Method	Prediction mean squared error (SD)	Improvement rate
Proposed ($\widehat{\boldsymbol{\beta}}$)	13.898 (4.427)	—
MRI Lasso	15.546 (5.715)	10.6%
PET Lasso	16.975 (7.009)	18.1%
Gene Lasso	19.946 (8.909)	30.3%
CC Lasso	19.956 (9.724)	30.4%
Naive mean	21.018 (10.410)	33.9%

7. Discussion

As mentioned in Section 2.1, methods that consider blockwise missing patterns, such as the proposed method and the method of Xue and Qu (2021), can incorporate both the complete case group and the incomplete groups in the imputation step, thus ensuring better accuracy. This is the main advantage of the proposed method compared with many existing imputation methods. However, our method may become complicated when there are too many data sources or different missing groups, leading to many blockwise imputations for each missing block, and thus many estimating equations to be solved. In general, the proposed method is more suitable for blockwise data with a small number of data sources and missing groups.

Although the MNAR mechanism is not covered in our theoretical justifications, simulation studies in Section 5 show that the proposed method still outperforms other methods under some MNAR settings. This may be because the proposed method incorporates more groups in the imputation of each missing block via the blockwise imputation. In this way, the proposed method aggregates information from various groups to reduce the selection bias in the groups caused by the MNAR mechanism. In future work, we may investigate managing MNAR situations by modeling the missingness or using instrumental variables.

A few other extensions are also worth exploring in the future. For example, because AD is a progressive brain disease, it is of interest to incorporate longitudinal data in the estimating functions to improve efficiency. In addition, currently, our method focuses only on linear regression with continuous responses; thus, it would be worthwhile generalizing our method to include binary and categorical responses.

Supplementary Material

We provide additional numerical and theoretical results and discussion, as well as proofs for all the theorems in the main text in the online Supplementary Material.

Acknowledgments

Fei Xue's research was partially supported by NSF Grant DMS-2210860. The authors thank the editor, associate editor, and anonymous reviewers for their helpful suggestions and comments. Additionally, this research was funded by NIH grant GM129781.

References

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology* **48**, 5–37.

Bartos, A., Gregus, D., Ibrahim, I. and Tintera, J. (2019). Brain volumes and their ratios in Alzheimer s disease on magnetic resonance imaging segmented using Freesurfer 6.0. *Psychiatry Research: Neuroimaging* **287**, 70–74.

Bellec, P. C., Dalalyan, A. S., Grappin, E. and Paris, Q. (2018). On the prediction loss of the lasso in the partially labeled setting. *Electronic Journal of Statistics* **12**, 3443–3472.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Cai, T., Cai, T. T. and Zhang, A. (2016). Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association* **111**, 621–633.

Cai, T. T. and Guo, Z. (2020). Semi-supervised inference for explained variance in high-dimensional regression and its applications. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **82**, 391–419.

Campos, S., Pizarro, L., Valle, C., Gray, K. R., Rueckert, D. and Allende, H. (2015). Evaluating imputation techniques for missing data in ADNI: A patient classification study. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 3–10. Springer International Publishing.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chandrasekher, K. A., Alaoui, A. E. and Montanari, A. (2020). Imputation for high-dimensional linear regression. *arXiv:2001.09180*.

Chapman, K. R., Bing-Canar, H., Alosco, M. L., Steinberg, E. G., Martin, B., Chaisson, C. et al. (2016). Mini mental state examination and logical memory scores for entry into Alzheimer's disease trials. *Alzheimer's Research & Therapy* **8**, 1–11.

Chen, L. S., Prentice, R. L. and Wang, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **70**, 312–322.

Deng, S., Ning, Y., Zhao, J. and Zhang, H. (2020). Optimal semi-supervised estimation and inference for high-dimensional linear regression. *arXiv:2011.14185*.

Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13**, 48–65.

Esteve, P., Rueda-Carrasco, J., Mateo, M. I., Martin-Bermejo, M. J., Draffin, J., Pereyra, G. et al. (2019). Elevated levels of Secreted-Frizzled-Related-Protein 1 contribute to Alzheimer's disease pathogenesis. *Nature Neuroscience* **22**, 1258–1268.

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

Garcia, R. I., Ibrahim, J. G. and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica* **20**, 149–165.

Gong, M., Ye, S., Li, W.-X., Zhang, J., Liu, Y., Zhu, J. et al. (2020). Regulatory function of praja ring finger ubiquitin ligase 2 mediated by the *P2rx3/P2rx7* axis in mouse hippocampal neuronal cells. *American Journal of Physiology-Cell Physiology* **318**, C1123–C1135.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**, 173–190.

Javanmard, A. and Javadi, H. (2019). False discovery rate control via debiased Lasso. *Electronic Journal of Statistics* **13**, 1212–1253.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15**, 2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics* **46**, 2593–2622.

Kang, K., Cai, J., Song, X. and Zhu, H. (2019). Bayesian hidden Markov models for delineating the pathology of Alzheimer's disease. *Statistical Methods in Medical Research* **28**, 2112–2124.

Kautzky, A., Seiger, R., Hahn, A., Fischer, P., Krampla, W., Kasper, S. et al. (2018). Prediction of autopsy verified neuropathological change of Alzheimer's disease using machine learning and MRI. *Frontiers in Aging Neuroscience* **10**. DOI: `10.3389/fnagi.2018.00406`.

Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics* **12**, 417–428.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–585.

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A. and Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: Biomarker analysis and shared morphometry database. *Scientific Reports* **8**, 1–16.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S. et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585.

Ma, R., Cai, T. T. and Li, H. (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association* **116**, 984–998.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W. et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics* **15**, 869–877.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research* **13**, 1665–1697.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical*

Science. A Review Journal of the Institute of Mathematical Statistics **27**, 538–557.

Neykov, M., Ning, Y., Liu, J. S. and Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science* **33**, 427–443.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

Qin, J., Zhang, B. and Leung, D. H. (2017). Efficient augmented inverse probability weighted estimation in missing data problems. *Journal of Business & Economic Statistics* **35**, 86–97.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11**, 2241–2259.

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**, 846–866.

Sanabria-Diaz, G., Martínez-Montes, E. and Melie-Garcia, L. (2013). Glucose metabolism during resting state reveals abnormal brain networks organization in the Alzheimer's disease and mild cognitive impairment. *PloS One* **8**, e68860.

Seaman, S. R. and Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. *Statistical Science* **33**, 184–197.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics* **6**, 38–90.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J. et al. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage* **102**, 192–206.

Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association* **116**, 1914–1927.

Yang, H., Xu, H., Li, Q., Jin, Y., Jiang, W., Wang, J. et al. (2019). Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *General Psychiatry* **32**, e100005.

Yu, G., Li, Q., Shen, D. and Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association* **115**, 1406–1419.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J. and Initiative, A. D. N. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage* **61**, 622–632.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 217–242.

Zhang, Y. and Bradic, J. (2019). High-dimensional semi-supervised learning: In search for optimal inference of the mean. *arXiv:1902.00772*.

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine* **4**, 9.

Zhou, Q., Goryawala, M., Cabrerizo, M., Barker, W., Duara, R. and Adjouadi, M. (2014). Significance of normalization on anatomical MRI measures in predicting Alzheimer's disease. *The Scientific World Journal* **2014**, 541802.

(Received October 2021; accepted May 2023)