Journal of Machine Learning Research 25 (2024) 1-36 Submitted 12/21; Revised 2/24; Published 2/24

Semi-supervised Inference for Block-wise Missing Data without Imputation

Shanshan Song SONGSS@TONGJI.EDU.CN
School of Mathematical Sciences and School of Economics and Management
Tongji University
Shanghai, 200092, China

Yuanyuan Lin[*] YLIN@STA.CUHK.EDU.HK
Department of Statistics
The Chinese University of Hong Kong
Hong Kong, China

Yong Zhou[*] YZHOU@FEM.ECNU.EDU.CN
Key Laboratory of Advanced Theory and Application in Statistics and Data Science, MOE,
Academy of Statistics and Interdisciplinary Sciences and School of Statistics
East China Normal University
Shanghai, 200062, China

Editor: Ji Zhu

Abstract

We consider statistical inference for single or low-dimensional parameters in a high-dimensional linear model under a semi-supervised setting, wherein the data are a combination of a labelled block-wise missing data set of a relatively small size and a large unlabelled data set. The proposed method utilises both labelled and unlabelled data without any imputation or removal of the missing observations. The asymptotic properties of the estimator are established under regularity conditions. Hypothesis testing for low-dimensional coefficients are also studied. Extensive simulations are conducted to examine the theoretical results. The method is evaluated on the Alzheimer's Disease Neuroimaging Initiative data.

Keywords: block-missing data, confidence intervals, hypothesis testing, semi-supervised inference

1. Introduction

Semi-supervised learning (Chapelle et al., 2009; Zhu and Goldberg, 2009) is a popular research field in statistics and machine learning because of the growing availability of unlabelled data and the costs of generating labelled data. Many semi-supervised data sets, which contain a small labelled data set and a large amount of unlabelled data, have been collected. Electronic health record (EHR) data sets are typically semi-supervised because labelling a subject with gold standard outcome is often costly and time-consuming. Earlier works on semi-supervised learning have focused on classification (Castelli and Cover, 1995, 1996; Blum and Mitchell, 1998; Nigam et al., 2000; Belkin and Niyogi, 2004; Weston et al., 2005; Wang and Shen, 2007; Wang et al., 2008), regression (Zhou and Li, 2005; Wasserman

[*] Corresponding Author

and Lafferty, 2007; Johnson and Zhang, 2008) and prediction (Liang et al., 2007; Ernst et al., 2008).

Recent advances in integrating unlabelled and labelled data for a combined analysis in a model-free framework have been reported in the literature. Chakrabortty and Cai (2018) proposed an efficient and adaptive semi-supervised estimator for linear regression. Gronsbell and Cai (2018) studied a class of semi-supervised approaches for the efficient evaluation of the predictive performance of logistic regression models. Zhang et al. (2019) studied semi-supervised estimation for the population mean of the response. The best linear approximation was examined by Azriel et al. (2021) under a semi-supervised setting. These important works have shown that semi-supervised estimators are more efficient than supervised estimators when the working model is mis-specified.

High-dimensional data are common in various scientific applications such as signal processing, econometrics and medical studies. State-of-the-art statistical methodologies have been developed for the inference of the regression coefficients in high-dimensional regression; see Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014), Ning and Liu (2017), Cai and Guo (2017), Javanmard and Montanari (2018), Belloni et al. (2019), among many others. By using unlabelled data, Bellec et al. (2018) provided non-asymptotic upper bounds for the prediction risk of the lasso estimator in the context of transductive and semi-supervised learning. Novel semi-supervised inference for the explained variance in a high-dimensional linear model was developed by Cai and Guo (2020). Deng et al. (2020) considered optimal semi-supervised inference in a working linear model.

Despite these developments, in many practical applications, the available data may be incompletely observed and subject to missingness among the high-dimensional covariates. The 'block missing' phenomenon frequently occurs when the data are from multiple sources or modalities. As different modalities may contain complementary information, statistical methods that use multi-modality data (rather than single-modality data) may yield better performance. In bioinformatics, a single measurement method is insufficient for examining the complex mechanisms of a disease. Nonetheless, it is often difficult to collect multiple measurements simultaneously for a single patient because (i) some measurements may be too expensive to be performed on each participant, and (ii) participants may be unwilling to take certain measurements because of physical or mental conditions. Generally, for multi-modality data, if the observations of a certain modality are completely missing, such data are termed *block-missing* data. Figure 1 shows three examples of block-missing data containing several modalities. For example, in Figure 1(a), the individuals in Groups 1-2 suffer from complete missingness of a certain modality. Interest has recently increased in developing statistical methodologies for analyzing block-missing data. Ignoring missing observations in statistical analysis is intuitive and easy-to-implement but unsurprisingly leads to substantial information loss (Nakagawa and Freckleton, 2008). Another promising approach to handling missing data is imputation (Xiang et al., 2014; Long and Johnson, 2015; Cai et al., 2016; Xue and Qu, 2020). Yu et al. (2020) studied a novel optimal sparse linear prediction for block-missing multi-modality data without imputation. In addition to the data subject to block missing among the covariates, Xue et al. (2021) developed an imputation-based semi-supervised inference procedure for a single coefficient in a high-dimensional linear model for which a large unlabelled data set is available.

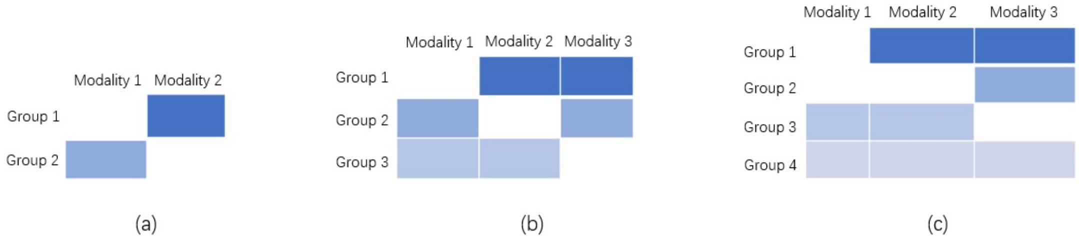

Figure 1: Different structures of block-missing data. Each blank block represents the missingness of a certain modality.

In this paper, without imputation, we propose a double debiased semi-supervised approach to conduct interval estimation and hypothesis testing for single or low-dimensional coefficients in a high-dimensional linear model with block-missing data. Our procedure consists of two steps: 1. To explore the association between the parameter of interest $\beta^\star$ and the partially-observed covariates, the ℓ_2 projection mapping the response on the space spanned by these observed covariates is considered, and the projection coefficient vector is denoted by θ. Building on the identified relationship between θ and $\beta^\star$, we design a extended lasso-type estimator (Tibshirani, 1996) of $\beta^\star$, where a debiasing step is applied to correct the biases incurred by the lasso estimator of θ. The resulting estimator is proven to satisfy the oracle inequality under mild conditions. 2. The Karush–Kuhn–Tucker (KKT) conditions corresponding to the designed minimization problem enable us to develop a bias-corrected estimator of $\beta^\star$, which has a Gaussian limiting distribution. Our proposed method avoids direct imputation, thus (i) it is more stable and computationally efficient, especially in high dimensions; (ii) no restrictive conditions are imposed on the association among the predictors; (iii) it is flexible to the block-missing structure in the sense that, no additional constraints are imposed on the index sets of observed covariates in different groups except for the constraint that $\cup_k \mathcal{I}(k) = \{1, \ldots, p\}$. Here $\mathcal{I}(k)$ denotes the index set of covariates that are observed in the k-th group. Notably, the nature of block-wise missing data results in different sample sizes among the components of the covariate vector; thus, the convergence rates of different components of our estimator for the regression parameter vector could be distinct but of the same order of $\mathcal{O}(\sqrt{n})$, where n is the size of the labelled data. This indicates that, unlike the lasso-type estimators whose convergence to the limit is not uniform, our bias-corrected estimator is not plagued by the nonuniformity of limiting distributions.

The rest of the paper is organized as follows. In Section 2, we describe the data representation, model assumption and the proposed semi-supervised inference procedure. A large-sample theory of the proposed estimator is presented in Section 3. In Section 4, the proposed method is evaluated by simulations and an analysis of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. A few concluding remarks are given in Section 5. All proofs of the theoretical results are presented in the Appendix.

2. Model and Methodology

2.1 Model and Data Description

Consider

$$Y = X^\top \beta^\star + \epsilon, \tag{1}$$

where Y is a scalar response variable and X is a p-vector of covariates, $\beta^\star \in \mathbb{R}^p$ is the true regression coefficient vector, and ϵ is an unobservable error term with mean 0 and variance σ^2, independent of X. Let $\beta_j^\star$ be the j-th component of $\beta^\star$, $j = 1, \dots, p$. Denote the joint distribution of (Y, X) as $\mathbb{P}_{Y,X}$ and the marginal distribution of X as $\mathbb{P}_X$. Let $\beta_G^\star \equiv \{\beta_j^\star : j \in G\}$, where G is any fixed-dimensional subset of $\{1, 2, \dots, p\}$. Denote $S_0 = \{j : \beta_j^\star \neq 0\}$ as the active set and its cardinality as $s_0 \equiv |S_0|$. Owing to the block-missing mechanism, we let K be the number of disjoint groups based on the missing patterns, and $\mathcal{S}_k$, $k = 1, \dots, K$, be the collection of individuals belonging to the k-th group. For instance, $K = 2$ in Figure 1(a).

Under a semi-supervised setting, the data available emanate from two sources: (i) $\mathcal{L} = \cup_{k=1}^K \mathcal{L}_k$, where $\mathcal{L}_k = \{(Y_i, Z_i^{(k)}), i \in \mathcal{S}_k\}$, $n_k = |\mathcal{S}_k|$ is the cardinality of $\mathcal{S}_k$, independent and identically distributed (i.i.d.) observations from the joint distribution of $(Y, Z^{(k)})$, where $\sum_k n_k = n$ and according to the nature of block-wise missingness, $Z^{(k)}$ is a certain sub-vector of X; (ii) $\mathcal{U} = \{X_i : i = n+1, \dots, n+N\}$ are N i.i.d. observations from $\mathbb{P}_X$. Let $\boldsymbol{X}^{(u)} \in \mathbb{R}^{N \times p}$ be the design matrix with rows $\{X_i : i = n+1, \dots, n+N\}$. Let $\mathcal{I}(k)$ be the index set of covariates that are observed in the k-th group. Assume that $\cup_k \mathcal{I}(k) = \{1, \dots, p\}$, that is, $X_{i\mathcal{I}(k)} = Z_i^{(k)}$ for any $i \in \mathcal{S}_k$ and $k = 1, \dots, K$. Let $p_k = |\mathcal{I}(k)|$ be the cardinality of $\mathcal{I}(k)$. For $j = 1, \dots, p$, let $\mathcal{H}(j)$ be the collection of groups with the j-th component of X observed. Throughout this paper, we assume that:

(A1) the commonly observed variables among different groups follow the same distribution $\mathbb{P}_{Y,X}$;

(A2) $\mathcal{L} \perp \mathcal{U}$, and $\mathcal{L}_k \perp \mathcal{L}_{k'}$ for $k \neq k'$, where "$\perp$" represents independence;

(A3) K is independent of (n, N), and $n/N \to 0$ as $n \to \infty$ and $N \to \infty$;

(A4) $p \to \infty$ as $n \to \infty$ and $N \to \infty$.

Remark 1 *When each modality is assumed to be missing completely at random, Condition (A1) holds. Condition (A1) also implies that the measured covariates from $\mathcal{L}$ and $\mathcal{U}$ follow the same distribution. In fact, it can be relaxed to Condition (A1'): for each $j \in \{1, \dots, p\}$, $k \in \mathcal{H}(j)$ and any $i \in \mathcal{S}_k$, the observed covariate X_{ij} satisfies that its first two moments, cross-covariance $\mathbb{E}\{X_{ij}X_{i\ell}\}$ for all $\ell \neq j$ with $\ell \in \mathcal{I}(k)$ and $\mathbb{E}\{X_{ij}Y_i\}$ are the same as those of the joint distribution $\mathbb{P}_{Y,X}$. Condition (A1') allows the commonly-observed covariates in $\mathcal{L}$ and $\mathcal{U}$ to follow different distributions, but those moment conditions in Condition (A1') should be satisfied.*

2.2 Double Debiased Semi-supervised Inference Procedure

Our goal is to conduct pointwise inference for a single component $\beta_j^\star$ or simultaneous inference for $\beta_G^\star$ using both labelled and unlabelled data. When the covariates X are completely observed in the labelled data set $\mathcal{L}$, since the lasso estimator (Tibshirani, 1996) is not root-n

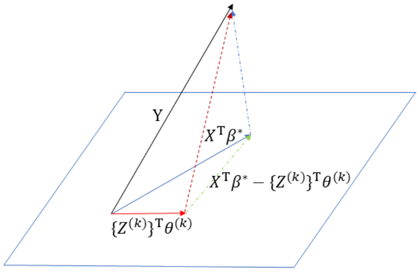
Y
$X^{\top}\beta^{*}$
$X^{\top}\beta^{*} - \{Z^{(k)}\}^{\top}\theta^{(k)}$
$\{Z^{(k)}\}^{\top}\theta^{(k)}$

Figure 2: Projection of $Y \in \mathbb{R}$ on the space of all linear combinations of $X \in \mathbb{R}^p$. Here, $Z^{(k)} \in \mathbb{R}^{p_k}$ is a subvector of X for $k = 1, \dots, K$.

consistent and does not have a tractable limiting distribution under the high-dimensional setting, effective debiasing methods were proposed (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014) for constructing valid confidence intervals and hypothesis testing under supervised setting. When the labeled data $\mathcal{L}$ suffer from block-missingness, Xue et al. (2021) developed a novel semi-supervised inference procedure with multiple blockwise imputation for each individual coefficient in a high dimensional linear model, where the missing values were imputed by fitting linear regression models with lasso. Such an imputation procedure has a computational complexity $\mathcal{O}((n+N)p^2 \min(n+N, p))$; see Section 2.12 in Bühlmann and van de Geer (2011) for more details. And its validity relies on the linearity assumption among covariates.

To avoid imputation, a natural idea is to explore the association between the target parameter $\beta^\star$ and the block-missing data $\mathcal{L}$ under model (1). For $k = 1, \dots, K$, we define $\theta^{(k)} = \arg\min_\theta \mathbb{E}([Y - \{Z^{(k)}\}^\top\theta]^2)$ and $\delta^{(k)} = Y - \{Z^{(k)}\}^\top\theta^{(k)}$, where the expectation "$\mathbb{E}$" is taken with respect to the joint distribution of $(Y, Z^{(k)})$. Then, $\theta^{(k)}$ is the ℓ_2 projection coefficient vector of Y onto $Z^{(k)}$, and $\{Z^{(k)}\}^\top\theta^{(k)}$ is the best linear predictor of Y given $Z^{(k)}$. For $k = 1, \dots, K$, $\theta^{(k)}$ is also the solution to $\mathbb{E}\left(Z^{(k)}\left[X^\top\beta^\star - \{Z^{(k)}\}^\top\theta^{(k)}\right]\right) = 0$, that is,

$$\mathbb{E}\left\{Z^{(k)}X^\top\beta^\star\right\} = \mathbb{E}\left[Z^{(k)}\{Z^{(k)}\}^\top\theta^{(k)}\right]. \tag{2}$$

Note that (2) always holds under model (1). We display the relationship between $\theta^{(k)}$ and $\beta^\star$ in Figure 2, indicating that $\delta^{(k)}$ can be decomposed into the sum of two orthogonal terms: ϵ and $X^\top\beta^\star - \{Z^{(k)}\}^\top\theta^{(k)}$, and $\theta^{(k)}$ can be also viewed as the ℓ_2 projection coefficient vector of $X^\top\beta^\star$ onto $Z^{(k)}$. To construct an estimating equation for estimating $\beta^\star$ based on (2), we need to estimate $\theta^{(k)}$ first. With n_k data points from $\mathcal{L}_k = \{(Y_i, Z_i^{(k)}), i \in \mathcal{S}_k\}$, the lasso estimator of $\theta^{(k)}$ is defined as

$$\hat{\theta}^{(k)} \in \arg\min_\theta \left(\frac{1}{n_k}\sum_{i\in\mathcal{S}_k}[Y_i - \{Z_i^{(k)}\}^\top\theta]^2 + \lambda_k\|\theta\|_1\right), \quad k = 1, \dots, K, \tag{3}$$

where λ_k is a tuning parameter. We use cross-validation to select λ_k among K groups in the numerical studies. In high dimensional setting, directly plugging in the regularized estimators $\hat{\theta}^{(k)}$ into (2) will result in inherent biases for estimating $\beta^\star$. Specifically, as stated in Cai et al. (2021), the biases can accumulate when projecting $\hat{\theta}^{(k)}$ along the direction of $Z_i^{(k)}$, which leads to a significant bias in estimating $\mathbb{E}[Z^{(k)}\{Z^{(k)}\}^\top\theta^{(k)}]$ and hence affects the convergence rate of the resulting estimator for $\beta^\star$.

To solve this problem, a debiasing step is needed to correct the biases incurred by $\hat{\theta}^{(k)}$. For each component of $\theta^{(k)}$, a bias-correction idea is to construct a projection direction by minimizing the variance with the bias constrained (Zhang and Zhang, 2014; Javanmard and Montanari, 2014). But it could be computationally expensive to identify such a projection direction for each component of $\theta^{(k)}$.

Our proposed debiasing idea is motivated by the error decomposition of the plug-in estimator $n_k^{-1}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}$:

$$\begin{aligned}\frac{1}{n_k}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\{Z_i^{(k)}\}^\top\hat{\theta}^{(k)} - \mathbb{E}\left[Z^{(k)}\{Z^{(k)}\}^\top\theta^{(k)}\right] &= -\frac{1}{n_k}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\left[Y_i - \{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}\right] \\ &\quad + \left(\frac{1}{n_k}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\{Z_i^{(k)}\}^\top - \mathbb{E}\left[Z^{(k)}\{Z^{(k)}\}^\top\right]\right)\theta^{(k)} + \frac{1}{n_k}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\delta_i^{(k)},\end{aligned} \tag{4}$$

where $\delta_i^{(k)} = Y_i - \{Z_i^{(k)}\}^\top\theta^{(k)}, i\in\mathcal{S}_k$. The first term $-n_k^{-1}\sum_{i=1}^{n_k} Z_i^{(k)}[Y_i-\{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}]$ on the right hand side of (4) is data-dependent, so the plug-in estimator $n_k^{-1}\sum_{i\in\mathcal{S}_k} Z_i^{(k)}\{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}$ subtracting $-n_k^{-1}\sum_{i=1}^{n_k} Z_i^{(k)}[Y_i - \{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}]$ shall be a better estimator for $\mathbb{E}[Z^{(k)}\{Z^{(k)}\}^\top \theta^{(k)}]$ with a faster convergence rate. Moreover, to estimate $\mathbb{E}[Z^{(k)}\{Z^{(k)}\}^\top]$, instead of using $n_k^{-1}\sum_{i=1} Z_i^{(k)}\{Z_i^{(k)}\}^\top$, we use both $\mathcal{L}_k$ and $\mathcal{U}$ to construct a new estimator:

$$\widetilde{\Sigma}_{n,N}^{(k)} = (N+n_k)^{-1}[\sum_{i=n+1}^{n+N} Z_i^{(k)}\{Z_i^{(k)}\}^\top + \sum_{i=1}^{n_k} Z_i^{(k)}\{Z_i^{(k)}\}^\top].$$

Hence, we propose the following calibrated semi-supervised estimator of $\mathbb{E}[Z^{(k)}\{Z^{(k)}\}^\top\theta^{(k)}]$:

$$SS_k = \widetilde{\Sigma}_{n,N}^{(k)}\hat{\theta}^{(k)} + \frac{1}{n_k}\sum_{i=1}^{n_k} Z_i^{(k)}\left[Y_i - \{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}\right].$$

Such a semi-supervised estimator is more accurate than the plug-in estimator when a large amount of unlabelled data are available. Our proposed debiasing approach is of independent interest and could be applied to other estimation problems.

Now, we are ready to introduce a point estimator for $\beta^\star$. Based on (2), we consider the following estimating equation:

$$\widehat{\boldsymbol{\Sigma}}_N\beta = \bar{S},$$

where $\widehat{\boldsymbol{\Sigma}}_N = N^{-1}\sum_{i=n+1}^{n+N} X_iX_i^\top$, $\bar{S} = (\bar{S}_1,\ldots,\bar{S}_p)^\top$ is a weighted average of $\{SS_k\}_{k=1}^K$ with $\bar{S}_j = [\sum_k \sqrt{n_k}I\{j\in\mathcal{I}(k)\}]^{-1}\sum_k\sqrt{n_k}SS_{k,(j)}I\{j\in\mathcal{I}(k)\}$, and $SS_{k,(j)}$ is the element of SS_k corresponding to the position of the j-th element of X in $Z^{(k)}$ if $j\in\mathcal{I}(k)$, $j=1,\ldots,p$. Heuristically, if $\widehat{\boldsymbol{\Sigma}}_N$ is invertible (which may not be true in a high-dimensional case), an estimator of $\beta^\star$ can be obtained by solving the above estimating equation, that is, $\widehat{\boldsymbol{\Sigma}}_N^{-1}\bar{S}$, which can be viewed as the minimizer of $\beta^\top \widehat{\mathbf{\Sigma}}_N \beta/2 - \bar{S}^\top \beta$. Thus, for large p, we consider an initial estimator of $\beta^\star$ defined as

$$\hat{\beta}_\lambda \in \arg\min_{\beta} \left(\frac{1}{2}\beta^\top \widehat{\mathbf{\Sigma}}_N \beta - \bar{S}^\top \beta + \lambda \|\beta\|_1 \right), \tag{5}$$

where λ is a tuning parameter. The selection of λ is discussed in Section 2.3, and some theoretical condition for λ is given in Theorem 2 in Section 3. The solution to the optimization problem in (5) might not be unique. Our theoretical results in Section 3 are valid for any minimizer of (5). In particular, under additional regularity conditions, we show that $\hat{\beta}_\lambda$ is consistent for $\beta^\star$ and also obtain its ℓ_1-norm estimation accuracy in Theorem 2. Although the initial estimator $\hat{\beta}_\lambda$ performs well in terms of point estimation, it is not root-n consistent and cannot be directly used for inference.

Next, we provide a new semi-supervised inference procedure for the low-dimensional coefficients $\beta^\star_G$, the main object of interest in our paper. Our main idea is to invert the Karush–Kuhn–Tucker (KKT) condition for problem (5) as in van de Geer et al. (2014), so that a test statistic taking the dependence among the components of the estimator of $\beta^\star_G$ into account can be constructed for simultaneous hypothesis testing problems for $\beta^\star_G$.

It follows from (2) and Lemma 9 in the Appendix A that for each $j = 1, 2, \ldots, p$,

$$\begin{aligned}
\bar{S}_j - \mathbb{E}(X_{ij} X_i^\top \beta^\star) &= \frac{1}{\sum_{k\in\mathcal{H}(j)} \sqrt{n_k}} \left[\sum_{k\in\mathcal{H}(j)} \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{S}_k} X_{ij} \left\{ Y_i - \{Z_i^{(k)}\}^\top \theta^{(k)} \right\} \right] + \Delta_{0,j} \\
&= \frac{1}{\sum_{k\in\mathcal{H}(j)} \sqrt{n_k}} \left\{ \sum_{k\in\mathcal{H}(j)} \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{S}_k} X_{ij} \delta_i^{(k)} \right\} + \Delta_{0,j},
\end{aligned} \tag{6}$$

where $\Delta_{0,j}$ is an asymptotically negligible term under mild conditions. Some notations are needed to present our idea conveniently. Let $\mathbf{X}_{fill}$ be the covariate matrix by filling the unobserved elements with 0. And let δ_w be the weighted residual vector composed by $\delta_i^{(k)}/\sqrt{n_k}, i \in \mathcal{S}_k$ from all groups as illustrated in Figure 3, which corresponds to the block-missing data structure in subfigure (c) in Figure 1. Let $\boldsymbol{J}$ be a $p \times p$ diagonal matrix with $\boldsymbol{J}_{j,j} = \sum_{k\in\mathcal{H}(j)} \sqrt{n_k}$, where $\mathcal{H}(j)$ denotes the collection of groups observing the j-th element of X in $\mathcal{L}$. Then, (6) can be recast into

$$\bar{S} - \mathbb{E}(XX^\top \beta^\star) = \boldsymbol{J}^{-1} \mathbf{X}_{fill}^\top \delta_w + \Delta_0,$$

where $\Delta_0 = (\Delta_{0,1}, \ldots, \Delta_{0,p})^\top$. It is known that $\hat{\beta}_\lambda$ defined in (5) satisfies the KKT conditions, that is, $-\left(\bar{S} - \widehat{\mathbf{\Sigma}}_N \hat{\beta}_\lambda\right) + \lambda\hat{\kappa} = 0$. Then,

$$\begin{aligned}
\hat{\beta}_\lambda - \beta^\star + \widehat{\mathbf{\Theta}}\lambda\hat{\kappa} &= \widehat{\mathbf{\Theta}} \boldsymbol{J}^{-1} \mathbf{X}_{fill}^\top \delta_w + \widehat{\mathbf{\Theta}}\Delta_0 - (\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}_N - \mathbf{I})(\hat{\beta}_\lambda - \beta^\star) - \widehat{\mathbf{\Theta}}(\widehat{\mathbf{\Sigma}}_N - \mathbf{\Sigma})\beta^\star \\
&\equiv \widehat{\mathbf{\Theta}} \boldsymbol{J}^{-1} \mathbf{X}_{fill}^\top \delta_w + \widehat{\mathbf{\Theta}}\Delta_0 + \Delta_1 + \Delta_2,
\end{aligned} \tag{7}$$

where $\widehat{\mathbf{\Theta}}$ is a proper approximate inverse of $\widehat{\mathbf{\Sigma}}_N$, $\Delta_1 = -(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}_N - \mathbf{I})(\hat{\beta}_\lambda - \beta^\star)$, $\Delta_2 = -\widehat{\mathbf{\Theta}}(\widehat{\mathbf{\Sigma}}_N - \mathbf{\Sigma})\beta^\star$ and $\mathbf{\Sigma} = \mathbb{E}(XX^\top)$. It is shown in Section 3 that $\widehat{\mathbf{\Theta}}\Delta_0$, Δ_1 and Δ_2 are

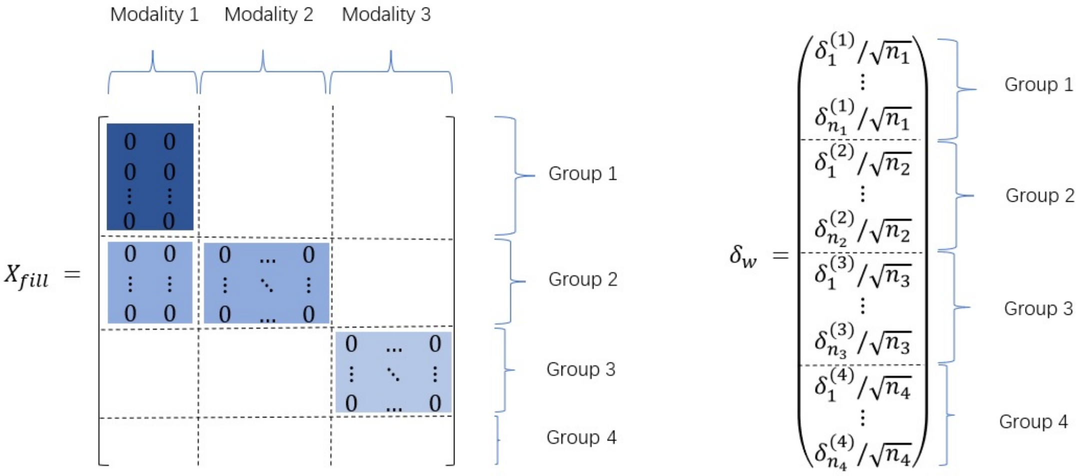
Modality 1
Modality 2
Modality 3
X_{fill} =
Group 1
Group 2
Group 3
Group 4
δ_w =
$\delta_1^{(1)}/\sqrt{n_1}$
$\delta_{n_1}^{(1)}/\sqrt{n_1}$
$\delta_1^{(2)}/\sqrt{n_2}$
$\delta_{n_2}^{(2)}/\sqrt{n_2}$
$\delta_1^{(3)}/\sqrt{n_3}$
$\delta_{n_3}^{(3)}/\sqrt{n_3}$
$\delta_1^{(4)}/\sqrt{n_4}$
$\delta_{n_4}^{(4)}/\sqrt{n_4}$
Group 1
Group 2
Group 3
Group 4

Figure 3: An illustration of $\mathbf{X}_{fill}$ and δ_w, corresponding to the block-missing structure of data presented in subfigure (c) in Figure 1.

asymptotically negligible under mild conditions. Thus, our proposed estimator of $\beta^\star$ is defined as

$$\hat{\beta} := \hat{\beta}_\lambda + \widehat{\mathbf{\Theta}}\left(\bar{S} - \widehat{\mathbf{\Sigma}}_N \hat{\beta}_\lambda\right).$$

Our method is essentially a double-debiasing procedure. The estimator $\hat{\beta}$ has a similar form to the de-sparsified lasso estimator introduced by van de Geer et al. (2014). In the following, we adopt the nodewise lasso (van de Geer et al., 2014) to construct $\widehat{\mathbf{\Theta}}$. Recall that $\boldsymbol{X}^{(u)}$ is the design matrix with rows $\{X_i : i = n+1, \ldots, n+N\}$. For each $j = 1, \ldots, p$, define $\hat{\gamma}_j = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \{\|\boldsymbol{X}^{(u)}_{\cdot j} - \boldsymbol{X}^{(u)}_{\cdot -j}\gamma\|_2^2/N + 2\lambda_j^{(u)}\|\gamma\|_1\}$ and $\hat{\tau}_j^2 = \|\boldsymbol{X}^{(u)}_{\cdot j} - \boldsymbol{X}^{(u)}_{\cdot -j}\hat{\gamma}_j\|_2^2/N + \lambda_j^{(u)}\|\hat{\gamma}_j\|_1$, where $\boldsymbol{X}^{(u)}_{\cdot j}$ is the j-th column of $\boldsymbol{X}^{(u)}$, $\boldsymbol{X}^{(u)}_{\cdot -j}$ is the submatrix of $\boldsymbol{X}^{(u)}$ after removing its j-th column and $\lambda_j^{(u)}$ is a tuning parameter. Rewrite $\hat{\gamma}_j = \{\hat{\gamma}_{j,k} : k = 1, \ldots, p, k \neq j\}$. Write $\widehat{D} = \operatorname{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2) \in \mathbb{R}^{p \times p}$. With a slight abuse of notations, we let $\widehat{\mathbf{\Theta}} = \widehat{D}^{-2}\widehat{C}$, where the j-th row of $\widehat{C} \in \mathbb{R}^{p \times p}$ is $\widehat{C}_j = (-\hat{\gamma}_{j,1}, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, -\hat{\gamma}_{j,p})^\top$, $j = 1, \ldots, p$.

Suppose that $\widehat{\mathbf{\Gamma}}$ is a component-wise consistent estimator of the limiting covariance matrix of $\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w$. For each $j = 1, \ldots, p$, an asymptotic $(1-\alpha)$-level confidence interval for $\beta_j^\star$ is given by

$$\left[\hat{\beta}_j - z_{\alpha/2}\sqrt{(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Theta}}^\top)_{j,j}}, \hat{\beta}_j + z_{\alpha/2}\sqrt{(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Theta}}^\top)_{j,j}}\right],$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0,1)$. The theoretical justifications for this procedure are provided in Section 3.

2.3 Tuning Parameter Selection

As the loss function in (5) cannot be expressed as an i.i.d. sum, the popular tuning parameter selectors, such as cross validation, AIC or BIC, are not directly applicable. Inspired by the extended Bayesian information criteria (EBIC) introduced by Chen and Chen (2008), we propose to minimize the following extended Bayesian information criterion for block-missing data (BM-EBIC) for selecting λ:

$$n_w \left(\hat{\beta}_\lambda^\top \widehat{\mathbf{\Sigma}}_N \hat{\beta}_\lambda - 2\bar{S}^\top \hat{\beta}_\lambda\right) + \hat{\sigma}_w^2 \|\hat{\beta}_\lambda\|_0 \log(n_w) + 2\hat{\sigma}_w^2 \log(\mathrm{C}_p^{\|\hat{\beta}_\lambda\|_0}), \tag{8}$$

where $n_w = p^{-1}\sum_{j=1}^{p}\sum_{k\in\mathcal{H}(j)} n_k^{3/2}/\boldsymbol{J}_{j,j}$ and $\hat{\sigma}_w^2 = p^{-1}\sum_{j=1}^{p}\sum_{k\in\mathcal{H}(j)} n_k(n_k - \|\hat{\theta}^k\|_0)^{-1} \sum_{i\in\mathcal{S}(k)}[Y_i - \{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}]^2/\boldsymbol{J}_{j,j}^2$ with $\boldsymbol{J}_{j,j} = \sum_{k\in\mathcal{H}(j)}\sqrt{n_k}$.

In view of the expression of n_w and $\hat{\sigma}_w^2$, the tuning parameter selector in (8) takes the block-missing structure of the data into account. Though our numerical studies in Section 4 contain supporting evidence that the proposed BM-EBIC selector works reasonably well, a rigorous proof of its consistency would be nontrivial and challenging, which merits further theoretical investigation.

2.4 Numerical Algorithm

We apply the proximal gradient descent with momentum algorithm to compute (5) numerically. The algorithm can be viewed as a combination of proximal gradient update and Nesterov's acceleration scheme (Nesterov, 2013). When the objective function contains a penalty term as in lasso, the algorithm is equivalent to the fast iterative soft-thresholding algorithm (FISTA) proposed by Beck and Teboulle (2009). The algorithm is summarized below. The soft thresholding operator $\mathcal{S}_\tau : \mathbb{R}^p \to \mathbb{R}^p$ with coordinates $(\mathcal{S}_\tau(a))_j = \mathrm{sign}(a_j)\max(|a_j| - \tau, 0)$ is used in the algorithm with threshold $\tau = s\lambda$. The step size is determined by the backtracking rule.

Algorithm 1 Proximal gradient descent with momentum

```
Require: some initialization $\beta^0$, $\alpha^0$, the tolerance parameter $tol$, backtracking rule parameters $(\tau, \gamma)$
Ensure: the solution in (5)
 1: $t \leftarrow 0$
 2: $D^0 \leftarrow 1$
 3: while $D^t > tol$ do
 4:     $\Delta^t = \widehat{\mathbf{\Sigma}}_N \beta^t - \bar{S}$                          ▷ Backtracking rule: steps 4-7
 5:     $s = 1$
 6:     while $f(\beta^t + s\Delta^t) > f(\beta^t) - \tau s\langle\Delta^t, \Delta^t\rangle$ do      ▷ Here, $f(\beta) = \beta^\top\widehat{\mathbf{\Sigma}}_N\beta/2 - \bar{S}^\top\beta$
 7:         $s \leftarrow \gamma s$
 8:     $\beta^{t+1} = \mathcal{S}_{s\lambda}\left(\alpha^t - s(\widehat{\mathbf{\Sigma}}_N\alpha^t - \bar{S})\right)$
 9:     $\alpha^{t+1} = \beta^{t+1} + \frac{t}{t+3}(\beta^{t+1} - \beta^t)$
10:     $D^{t+1} = \|\beta^{t+1} - \beta^t\|$
11:     $t \leftarrow t + 1$
```

One may also consider the majorize–minorize (MM) algorithm, another important alternative method to solve penalized estimating equations (Johnson et al., 2008).

3. Asymptotic Properties

More notations are needed. For a number $c \in \mathbb{R}$, c_p denotes a p-vector with each element being c. For an index set $S \subseteq \{1, \dots, p\}$, a_S is the subvector of $a \in \mathbb{R}^p$ consisting of all components a_j with $j \in S$. Denote a_{-j} as the subvector of a after removing the j-th element. Let $|S|$ be the cardinality of a set S. For a matrix $\boldsymbol{A}$, we define $\|\boldsymbol{A}\|_\infty = \max_{i,j} |\boldsymbol{A}_{i,j}|$. For index sets $S, S_1, S_2 \subseteq \{1, \dots, p\}$, $\boldsymbol{A}_{S\cdot}$ and $\boldsymbol{A}_{\cdot S}$ represent the submatrix of $\boldsymbol{A}$ consisting of its rows and columns indexed by S, respectively, and $\boldsymbol{A}_{S_1,S_2}$ denotes the submatrix of $\boldsymbol{A}$ consisting of entries in the rows indexed by S_1 and the columns indexed by S_2. $C, C_1, C_2, \dots$ are generic constants that may vary from place to place. For two sequences of real numbers a_n and b_n, $a_n = o(b_n)$ if $a_n/b_n \to 0$; $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a constant C such that $a_n \le C b_n$ for all n. Denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

The following Conditions (B1)-(B3) on the underlying model and Conditions (C1)-(C3) on the block-missing mechanism are imposed.

(B1) The covariate vector X follows zero-mean sub-Gaussian distribution.

(B2) The smallest eigenvalue Λ_{min} of $\boldsymbol{\Sigma}$ is bounded by a positive constant C_1, i.e., $\Lambda_{min} \ge C_1$.

(B3) The second moment of Y is finite, i.e., $\mathbb{E}(Y^2) < \infty$.

(C1) The number of groups K is finite.

(C2) For $k = 1, \dots, K$, the error term $\delta^{(k)}$ follows sub-Gaussian distribution with mean 0 and variance $\{\eta^{(k)}\}^2 < \infty$.

(C3) Assume $\max_{1\le k\le K} s^{(k)}\sqrt{\log p_k/n_k} = o(1)$, where $s^{(k)} = \|\theta^{(k)}\|_0$.

These conditions are regularity conditions. When the common distribution assumption in Condition (A1) is relaxed as Condition (A1') in Remark 1, our theoretical results in this section will still hold if Condition (C1) is replaced by Condition (C1'): There exist positive constants C, ν such that for each $k = 1, \dots, K$, any $j \in \mathcal{I}(k)$, any $i \in \mathcal{S}_k$ and every $t > 0$, X_{ij} is a zero-mean random variable satisfying

$$\mathbb{P}(|X_{ij}| > t) \le C \exp^{-\nu t^2},$$

where $\mathcal{I}(k)$ denotes the index set of covariates observed in the k-th group. Condition (C3) is the sparsity condition for high-dimensional models. More discussions on the sparsity of $\theta^{(k)}$ can be found in Bühlmann and van de Geer (2015).

The cardinality of the active set $S_0 = \{j : \beta_j^\star \ne 0\}$ is denoted by s_0 and $s^{(k)} = \|\theta^{(k)}\|_0$. For ease of exposition, we denote

$$\tau(p_k, s^{(k)}, n_k, N) = \left\{ \frac{N}{(N+n_k)}\sqrt{\frac{\log p_k}{n_k}} + \frac{\sqrt{N\log p_k}}{(N+n_k)} \right\} s^{(k)}\sqrt{\frac{\log p_k}{n_k}} + \sqrt{\frac{s^{(k)}\log p_k}{(N+n_k)}}.$$

Theorem 2 *(Oracle inequality) Suppose that Conditions (A1)-(A4), (B1)-(B3), (C1)-(C3) hold and* $\log p_k \lesssim \min\{N, n_k\}, \log p \lesssim \min\{N, n_1, \dots, n_K\}$. *Then, for* $\lambda_k \asymp \sqrt{\log p_k/n_k}$, $k = 1, \dots, K$ *and* $\lambda \asymp \max_j[\{\sum_{k\in\mathcal{H}(j)} \sqrt{n_k}\}^{-1}|\mathcal{H}(j)|\sqrt{\log p}] + \sqrt{s_0 \log p/N} + \max_j[\{\sum_{k\in\mathcal{H}(j)}$

$\sqrt{n_k}\}^{-1}\sum_{k\in\mathcal{H}(j)}\{\sqrt{n_k}\tau(p_k,s^{(k)},n_k,N)\}]$ *and* $s_0 \ll \sqrt{N/\log p}$, *the event*

$$(\hat{\beta}_\lambda-\beta^\star)^\top\widehat{\boldsymbol{\Sigma}}_N(\hat{\beta}_\lambda-\beta^\star)+\lambda\|\hat{\beta}_\lambda-\beta^\star\|_1 \lesssim \lambda^2 s_0, \tag{9}$$

holds with the probability larger than $1-2p^{-C_1}-10K\{\min_k p_k\}^{-C_2}$ *for absolute constants* $C_1>0, C_2>0$.

Theorem 2 gives an oracle inequality for the initial estimator $\hat{\beta}_\lambda$ in (5). For a new pair of observation (Y,X), the conditional prediction error $\mathbb{E}\{(Y-X^\top\hat{\beta}_\lambda)^2|\mathcal{L},\mathcal{U}\}=\sigma^2+(\hat{\beta}_\lambda-\beta^\star)^\top\boldsymbol{\Sigma}(\hat{\beta}_\lambda-\beta^\star)$. Theorem 2 implies that $(\hat{\beta}_\lambda-\beta^\star)^\top\widehat{\boldsymbol{\Sigma}}_N(\hat{\beta}_\lambda-\beta^\star)\lesssim\lambda^2 s_0$ holds with high probability. In addition, Theorem 2 also gives the bound for the ℓ_1-error of $\hat{\beta}_\lambda$, i.e., $\|\hat{\beta}_\lambda-\beta^\star\|_1\lesssim\lambda s_0$ holds with high probability.

More remarks on Theorem 2 can be made. First, if we replace the proposed estimator SS_k by the plug-in estimator $n_k^{-1}\sum_{i=1}^{n_k}Z_i^{(k)}\{Z_i^{(k)}\}^\top\hat{\theta}^{(k)}$ in the construction of the initial estimator $\hat{\beta}_\lambda$, Theorem 2 still holds if $\lambda \asymp \max_j[\{\sum_{k\in\mathcal{H}(j)}\sqrt{n_k}\}^{-1}|\mathcal{H}(j)|\sqrt{\log p}]+\sqrt{s_0\log p/N}+\max_j[\{\sum_{k\in\mathcal{H}(j)}\sqrt{n_k}\}^{-1}\sum_{k\in\mathcal{H}(j)}\{\ \sqrt{s^{(k)}\log p_k}\}]$. The key difference is in the two terms $\tau(p_k,s^{(k)},n_k,N)$ and $\sqrt{s^{(k)}\log p_k/n_k}$. Since the order of $\tau(p_k,s^{(k)},n_k,N)$ is much smaller than that of $\sqrt{s^{(k)}\log p_k/n_k}$, our proposed initial estimator $\hat{\beta}_\lambda$ has a much faster convergence rate. Especially, when $N\gtrsim\max_k[n^2/\{s^{(k)}\log p_k\}]$, $\tau(p_k,s^{(k)},n_k,N)$ is of order $s^{(k)}\log p_k/n_k$. Second, the oracle inequality in (9) also shows how the unlabelled data contribute to the convergence rate of $\hat{\beta}_\lambda$. To explain this, we consider three special cases when $p_1\asymp p+_2\asymp\ldots\asymp p_K\asymp p$ and $s^{(1)}\asymp s^{(2)}\asymp\ldots\asymp s^{(K)}\asymp s_0$.

- Case (i): when $n_1\asymp n_2\asymp\ldots\asymp n_K\asymp n$, we have

  $$\|\hat{\beta}_\lambda-\beta^\star\|_1\lesssim s_0\sqrt{\log p/n}+s_0^{3/2}\sqrt{\log p/N}+(N+n)^{-1}Ns_0^2(\log p/n+\log p/\sqrt{nN}), \tag{10}$$

  holds with high probability. Recall that with n labelled samples, the convergence rate of the ℓ_1-error of the classical lasso estimator is $\mathcal{O}(s_0\sqrt{\log p/n})$, where $s_0\ll\sqrt{n/\log p}$ (Bühlmann and van de Geer, 2011). Hence, the addtional term $s_0^{3/2}\sqrt{\log p/N}+(N+n)^{-1}Ns_0^2(\log p/n\ +\log p/\sqrt{nN})$ in (10) is the price paid for the block-missingness of labelled data. In this case, if the unlabelled data size satisfies $N\gtrsim s_0 n$, then $\|\hat{\beta}_\lambda-\beta^\star\|_1\lesssim s_0\sqrt{\log p/n}$ holds with high probability, i.e., the convergence rate of the ℓ_1-error of $\hat{\beta}_\lambda$ attains $s_0\sqrt{\log p/n}$.

- Case (ii): when $n\asymp n_1\asymp n_2\asymp\ldots\asymp n_{K-1}\gg n_K$ with all complete samples coming from the K-th group (i.e., with n_K completely-observed samples available),

  $$\|\hat{\beta}_\lambda-\beta^\star\|_1\lesssim s_0\sqrt{\log p/n}+s_0^{3/2}\sqrt{\log p/N}+(N+n_K)^{-1}Ns_0^2(\log p/\sqrt{n_K n}+\log p/\sqrt{nN}),$$

  with high probability. In this case,

  - if $N\asymp s_0 n_K$, $\|\hat{\beta}_\lambda-\beta^\star\|_1\lesssim s_0\sqrt{\log p/n_K}$ holds with high probability, which is at the same order as the ℓ_1-error of the lasso estimator based on n_K complete observations only;

  - if $s_0 n_K \lesssim N \ll s_0 n$, $\|\hat{\beta}_\lambda - \beta^\star\|_1 \lesssim s_0^{3/2}\sqrt{\log p/N}$ holds with high probability, indicating that our initial estimator $\hat{\beta}_\lambda$ has a faster convergence rate of the ℓ_1-error than the lasso estimator based on n_K complete observations only;
  - if $N \gtrsim s_0 n$, $\|\hat{\beta}_\lambda - \beta^\star\|_1 \lesssim s_0\sqrt{\log p/n}$ holds with high probability, implying that the initial estimator $\hat{\beta}_\lambda$ achieves the optimal convergence rate $\mathcal{O}(s_0\sqrt{\log p/n})$ if the unlabelled data size is sufficiently large.

- Case (iii): when $n_1 \asymp n_2 \asymp \ldots \asymp n_{K-1} \ll n_K \asymp n$ with all complete samples coming from the K-th group, if $s_0 \ll \sqrt{n_1/\log p}$, we have

$$\|\hat{\beta}_\lambda - \beta^\star\|_1 \lesssim s_0\sqrt{\log p/n} + s_0^{3/2}\sqrt{\log p/N} + (N+n_1)^{-1}Ns_0^2(\log p/\sqrt{n_1 n} + \log p/\sqrt{nN}),$$

  holds with high probability. Moreover, when the unlabelled data size $N \gtrsim s_0 n$, $\|\hat{\beta}_\lambda - \beta^\star\|_1 \lesssim s_0\sqrt{\log p/n}$ holds with high probability.

Several stimulating works on semi-supervised inference (Chakrabortty and Cai, 2018; Deng et al., 2020; Azriel et al., 2021) have shown that, when the linear model is correctly specified, improved estimation of the regression coefficients using additional unlabelled data is impossible without further assumptions relating the target parameters to the marginal distribution of X. This is indeed true when all labelled data are completely observed. Our Theorem 2 shows that the unlabelled data play an important role in improving the convergence rate of the initial estimator in the presence of block-missing labelled data.

Under some mild conditions, it was shown in Theorem 1 in Yu et al. (2020) that the convergence rate of their sparse estimator for the coefficients in the optimal linear prediction in terms of ℓ_2-norm is bounded by $\sqrt{s_0 \log p}\{\min_j \sum_{k\in\mathcal{H}(j)} n_k\}^{-1/2}$ under our notations, where $\mathcal{H}(j)$ is the collection of groups with the j-th element of X observed. When the covariates in $\mathcal{L}$ and $\mathcal{U}$ are blockwise missing and the missing patterns are shared between $\mathcal{L}$ and $\mathcal{U}$, Xue et al. (2021) proposed an imputation-based estimator for $\beta^\star$ and proved that the convergence rate of their estimator under ℓ_2-norm is at the order of $\sqrt{s_0 \log p/n}$ when the size of unlabelled data $N \gtrsim s_0^2 n$ under some conditions.

In the next theorem, we will show how the unlabelled data contribute to the convergence rate of the proposed estimator $\hat{\beta}$. Define $\gamma_j = \arg\min_{\gamma\in\mathbb{R}^{p-1}} \mathbb{E}\{X_{ij} - X_{i-j}^\top\gamma\}^2$ and $\tau_j^2 = \mathbb{E}[\{X_{ij} - X_{i-j}^\top\gamma_j\}^2]$ for $j = 1,\ldots,p$, where X_{i-j} is the covariate subvector after removing the j-th element of X_i. Let $\mathbf{\Theta} \equiv \mathbf{\Sigma}^{-1}$ and define $s_j = |\{k \neq j : \mathbf{\Theta}_{j,k} \neq 0\}|$. Additional assumptions are needed.

(D1) Assume that $\max_{1\le j\le p} s_j\sqrt{\log p/N} = o(1)$.

(D2) There exists a positive constant C_2 such that $\max_j \mathbf{\Sigma}_{j,j} \le C_2$.

(D3) Assume $\max_{1\le k\le K} s^{(k)}\log p_k/\sqrt{n_k} = o(1)$, where $s^{(k)} = \|\theta^{(k)}\|_0$.

Conditions (D1)-(D2) are to ensure well behavior of $\widehat{\mathbf{\Theta}}$. Condition (D3) is a regular condition on the sparsity of $\theta^{(k)}$.

Theorem 3 *Suppose that Conditions (A1)-(A4), (B1)-(B2), (D1)-(D2) hold. Then, if $\lambda_j^{(u)} \asymp \sqrt{\log p/N}$ uniformly in $j \in \{1, 2, \dots, p\}$ for the nodewise lasso, we have*

$$\|\widehat{\mathbf{\Theta}}_{j\cdot} - \mathbf{\Theta}_{j\cdot}\|_1 = O_p\left(s_j\sqrt{\frac{\log p}{N}}\right), \qquad \|\widehat{\mathbf{\Theta}}_{j\cdot} - \mathbf{\Theta}_{j\cdot}\|_2 = O_p\left(\sqrt{\frac{s_j \log p}{N}}\right),$$
$$|\hat{\tau}_j^2 - \tau_j^2| = O_p\left(\sqrt{\frac{s_j \log p}{N}}\right),$$

uniformly in $j \in \{1, 2, \dots, p\}$, where $\widehat{\mathbf{\Theta}}_{j\cdot}$ is the j-th row of $\widehat{\mathbf{\Theta}}$. Additionally, if Conditions (B3), (C1)-(C2) and (D3) hold, $n_1 \asymp n_2 \asymp \dots \asymp n_K \asymp n$, $s_0 \ll \min\{N/n, \sqrt{N}/\log p, N/(\max_j s_j n \log p)\}$ for $N \gtrsim \max_k[n_k^2/\{s^{(k)} \log p_k\}]$, $\max_j \sqrt{s_j}\ \max_k\{s^{(k)} \log p_k/\sqrt{n}\} = o(1)$, then, for $\lambda_k \asymp \sqrt{\log p_k/n}$, $k = 1, \dots, K$ in (3) and $\lambda \asymp \sqrt{\log p/n}$ in (5), we have

$$\sqrt{n}(\hat{\beta} - \beta^\star) = \sqrt{n}\widehat{\mathbf{\Theta}}\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w + \Delta, \tag{11}$$

where $\|\Delta\|_\infty = o_p(1)$, as $n \to \infty$ and $N \to \infty$.

The proof of Theorem 3 is given in the Appendix B. The proof of (11) is nontrivial, as we need to show that $\widehat{\mathbf{\Theta}}\Delta_0$, Δ_1 and Δ_2 in (7) are asymptotically negligible, i.e., $\sqrt{n}\|\widehat{\mathbf{\Theta}}\Delta_0\|_\infty = o_p(1)$, $\sqrt{n}\|\Delta_1\|_\infty = o_p(1)$ and $\sqrt{n}\|\Delta_2\|_\infty = o_p(1)$ under block-wise missing setting. In contrast, if all labelled data are completely observed, the two terms $\sqrt{n}\|\widehat{\mathbf{\Theta}}\Delta_0\|_\infty$ and $\sqrt{n}\|\Delta_2\|_\infty$ are not involved; and van de Geer et al. (2014) showed that $\sqrt{n}\|\Delta_1\|_\infty = o_p(1)$ under the assumption that $s_0 \ll \sqrt{n}/\log p$ and $\max_j s_j \log p/n = o(1)$. Hence, those additional conditions for guaranteeing $\sqrt{n}\|\widehat{\mathbf{\Theta}}\Delta_0\|_\infty = o_p(1)$ and $\sqrt{n}\|\Delta_2\|_\infty = o_p(1)$ can be viewed as the price paid under block-wise missing setting. In addition, $s_0 \ll \min\{N/n, \sqrt{N}/\log p, N/(\max_j s_j n \log p)\}$ indicates that the requirement on the sparsity of $\beta^\star$ is loose if N is large.

Theorem 3 has some important implications. As the labelled data $\mathcal{L}$ and the unlabelled data $\mathcal{U}$ are independent, $\widehat{\mathbf{\Theta}}$ is independent of $\sqrt{n}\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w$, which differs from that in van de Geer et al. (2014). However, recall that $\boldsymbol{J}$ is a $p \times p$ diagonal matrix with $\boldsymbol{J}_{j,j} = \sum_{k \in \mathcal{H}(j)} \sqrt{n_k}$, where $\mathcal{H}(j)$ denotes the collection of groups with the j-th element of X in $\mathcal{L}$ observed. The diagonal elements of $\boldsymbol{J}^{-1}$ are generally distinct, resulting in the different convergence rates of $\hat{\beta}_j$ among different $j = 1, \dots, p$ in practice, albeit at the same order $\mathcal{O}(\sqrt{n})$ when $n_1 \asymp n_2 \asymp \dots \asymp n_K \asymp n$. This is due to the block-missing structure of the data. Moreover, unlike the initial estimator $\hat{\beta}_\lambda$, the proposed estimator $\hat{\beta}$ does not suffer from those problems due to the nonuniformity of limiting theory. Based on Theorem 3, for each $j = 1, \dots, p$, $\sqrt{n}(\hat{\beta}_j - \beta_j^\star)$ is asymptotically normally distributed. We can then construct a pointwise confidence interval for $\beta^\star$. The next corollary is a direct consequence of Theorem 3 which holds in a uniform sense.

Corollary 4 *Under the conditions given in Theorem 3, for any fixed-dimensional subset $G \subseteq \{1, \dots, p\}$, we have*

$$\hat{\beta}_G - \beta_G^\star = \widehat{\mathbf{\Theta}}_{G\cdot}\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w + n^{-1/2}\Delta_G,$$

where $\|\Delta_G\|_\infty = o_p(1)$ as $n \to \infty$ and $N \to \infty$.

Corollary 4 provides a theoretical basis for constructing simultaneous confidence regions for $\beta^\star_G$. Recall that an estimator $\widehat{\mathbf{\Gamma}}$ of the limiting covariance matrix of $\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w$ is needed as described in Section 2.2 in the construction of the pointwise confidence interval of $\beta^\star$. We construct an estimator $\widehat{\mathbf{\Gamma}}$ in the following. For $k = 1, \dots, K$, if $j \in \mathcal{I}(k)$, denote $\omega_{j,j}^{(k)} = \mathbb{E}\{|X_{ij}\delta_i^{(k)}|^2\}$ for $i \in \mathcal{S}_k$. We propose to estimate $\omega_{j,j}^{(k)}$ by

$$\hat{\omega}_{j,j}^{(k)} := \frac{1}{n_k - \hat{s}_k}\sum_{i\in\mathcal{S}_k}\left\{\hat{\delta}_i^{(k)}X_{ij} - \frac{1}{n_k}\sum_{i\in\mathcal{S}_k}\hat{\delta}_i^{(k)}X_{ij}\right\}^2, \quad \hat{\delta}_i^{(k)} = Y_i - \{Z_i^{(k)}\}^\top\hat{\theta}_k,$$

where $\hat{s}_k = \|\hat{\theta}_k\|_0$. For $j, j' \in \mathcal{I}(k)$, a similar estimator of $\omega_{j,j'}^{(k)} = \mathbb{E}[X_{ij}X_{ij'}\{\delta_i^{(k)}\}^2]$ can be constructed, denoted by $\hat{\omega}_{j,j'}^{(k)}$. Then, the set $\{\hat{\omega}_{j,j'}^{(k)}/n_k : j, j' \in \mathcal{I}(k), k = 1, \dots, K\}$ is used to construct $\widehat{\mathbf{\Gamma}}$. The next theorem conveys that the resulting estimator is an element-wise consistent estimator for the limiting covariance matrix of $\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w$, which can be served as $\widehat{\mathbf{\Gamma}}$ in practice.

Theorem 5 *Suppose that Conditions (A1)-(A4), (B1)-(B2), (C2) and (D3) hold. Then,*

$$\hat{\omega}_{j,j}^{(k)}/\omega_{j,j}^{(k)} = 1 + o_p(1).$$

Hypothesis testing for single or low-dimensional coefficients in a high-dimensional linear model is another inference problem. Recall that $\beta^\star_G = \{\beta^\star_j : j \in G\}$, where $G \subseteq \{1, 2, \dots, p\}$ is any subset of a fixed dimension. One can construct a statistical test for the null hypothesis $H_0 : \beta^\star_G = 0$ using the test statistic

$$T := \left\|(\widehat{\mathbf{\Theta}}_{G\cdot}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Theta}}_{G\cdot}^\top)^{-1/2}\hat{\beta}_G\right\|_2^2. \tag{12}$$

Corollary 4 implies that under the null hypothesis H_0, the limiting distribution of the test statistic T is $\chi^2(|G|)$, the chi-squared distribution with degree of freedom $|G|$. One may reject H_0 if $T > \chi^2_{1-\alpha}(|G|)$, where $\chi^2_{1-\alpha}(|G|)$ is the lower $(1-\alpha)$-quantile of $\chi^2(|G|)$. This test statistic takes the dependence among the components of $\hat{\beta}_G$ into account.

Our proposed semi-supervised inference procedure is imputation-free, thus (1) it is computationally efficient, especially in high dimensions; (2) no restrictive conditions are imposed on the correlation among the predictors X; (3) it is flexible to the block-missing mechanism of data in the sense that no additional constraints are imposed on the relationship between the index sets of observed covariates in different groups except the constraint that $\cup_k\mathcal{I}(k) = \{1, \dots, p\}$, where $\mathcal{I}(k)$ denotes the index set of covariates that are observed in the k-th group. Thus, our method works for the scenarios in Figure 1(a). Moreover, our proposed confidence interval does not require complete labelled observations and it is most useful when the proportion of complete labelled observations is relatively low or even zero. Nonetheless, with the availability of a relatively large size of complete labelled observations, our method may not be asymptotically as efficient as the debiasing estimators studied by van de Geer et al. (2014) and Zhang and Zhang (2014) which use completely-observed labelled samples only, albeit at the same order of the convergence rate.

4. Numerical Studies

4.1 Simulation Studies

We investigate the finite-sample performance of our proposed method under several scenarios. For any set $S \subseteq \{1, \dots, p\}$, the average empirical coverage probability (ACP) and average length (AL) of 95% confidence intervals over S are defined as

$$\mathrm{ACP}(S) = \sum_{j \in S} \mathrm{CP}_j / |S|, \qquad \mathrm{AL}(S) = \sum_{j \in S} \mathrm{CIL}_j / |S|,$$

where CP_j and CIL_j represent the empirical coverage probability and the length of the 95% confidence interval for $\beta_j^\star$, respectively. For any fixed-dimensional subset $G \subseteq \{1, 2, \dots, p\}$, we consider to test $H_0 : \beta_G^\star = 0$. If G is an active subset, i.e., H_0 is false, we report the empirical power of our test statistic T defined in (12); if G is an inactive subset, i.e., H_0 is true, we report the empirical size of the test statistic T:

$$\mathrm{Power} = \sum_{r=1}^{R} I\{T_r > \chi^2_{1-\alpha}(|G|)\}, \qquad \mathrm{Size} = \sum_{r=1}^{R} I\{T_r > \chi^2_{1-\alpha}(|G|)\},$$

where R is the number of replications, T_r denotes the value of the test statistic in the r-th replication, $r = 1, \dots, R$, and $\chi^2_{1-\alpha}(|G|)$ is the lower $(1-\alpha)$-quantile of $\chi^2(|G|)$. We set $\alpha = 0.05$. The empirical power is the proportion of rejecting a false null hypothesis, and the empirical size is the proportion of rejecting a true null hypothesis. The larger Power is, the better the proposed test is; and the Size is close to the nominal significance level, implying that the test is valid.

Three error distributions are tried: (1) Standard normal distribution, denoted by $N(0, 1)$; (2) Student's t distribution with degrees of freedom 3, denoted by $t(3)$; (3) Weibull distribution with shape parameter 0.5 and scale parameter 0.3, denoted by $WB(0.5, 0.3)$. We consider three simulated examples below:

- (**E1**): The predictor vector X follows Gaussian distribution $N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = 0.4^{|i-j|}$, and the three distributions of ϵ are tried. We set the target parameter $\beta^\star = (0.8_3, 0_{p/2-3}, 0.8_3, 0_{p/2-3})^\top$, where p is the dimensionality of X. The number of covariates in the active set $s_0 = 6$. We consider the block-missing structure with 2 modalities, as in Figure 1(a). The labelled samples are uniformly assigned to the two groups and the unlabelled data are independently generated from $N(0, \mathbf{\Sigma})$. Here, $K = 2$, $n = 200$ or 300, $p = 450$, $N = 1000$ or 5000, $p_1 = p_2 = 225$ and $n_1 = n_2 = n/2$.

- (**E2**): The predictor vector X follows (i) Gaussian distribution $N(0, \mathbf{\Sigma})$ with the covariance matrix satisfying $\mathbf{\Sigma}_{i,j} = 0.4^{|i-j|}$ or (ii) the Gaussian mixture distribution $0.5N(0, \mathbf{\Sigma}^{(1)}) + 0.5N(0, \mathbf{\Sigma}^{(2)})$ with $\mathbf{\Sigma}^{(1)}_{i,j} = 0.4^{|i-j|}$, $\mathbf{\Sigma}^{(2)}_{i,j} = 0.4^{2|i-j|}$ and the above three settings of ϵ are considered. We set the target parameter $\beta^\star = (0.8_3, 0_{p/3-3}, 0.8_3, 0_{p/3-3}, 0.8_3, 0_{p/3-3})^\top$ and $s_0 = 9$. The block-missing structure with 3 modalities is presented in Figure 1(b). The Labelled samples are uniformly assigned to the three groups and the unlabelled data are independently generated from (i) or (ii). Here, $K = 3$, $n = 300$ or 600, $p = 150$, $N = 5000$, $p_1 = p_2 = p_3 = 50$ and $n_1 = n_2 = n_3 = n/3$.

- (**E3**): The predictor vector X follows Gaussian distribution $N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = 0.4^{|i-j|}$ and $\epsilon \sim N(0,1)$. Set the target parameter $\beta^\star = (0.8_3, 0_{p/3-3}, 0.8_3, 0_{p/3-3}, 0.8_3, 0_{p/3-3})^\top$ and $s_0 = 9$. We consider the block-missing structure with 3 modalities, as in Figure 1(c). Samples with incomplete observations are randomly assigned to the first three groups with probabilities $(0.4, 0.3, 0.3)$; samples with complete observations are generated independently; the unlabelled data are independently generated from $N(0, \mathbf{\Sigma})$. Here, $K = 4$, $n = 360, 400$ or 500, $p = 150$, $N = 5000$ or 5000, $p_1 = 30, p_2 = 90, p_3 = 60, p_4 = 150$ and $n_4 = 60, 100$ or 200.

For pointwise confidence intervals and hypothesis testing problems, we compare our proposed method with, if applicable, (i) Z. & Z.: the debiasing method proposed by Zhang and Zhang (2014) with scaled lasso using complete observations only; (ii) G. et al.: the debiasing method by van de Geer et al. (2014) with lasso using complete observations only; (iii) X. et al.: a revised version of the imputation-based method studied by Xue et al. (2021). Since Xue et al. (2021) focused on the cases that the covariates of the samples in $\mathcal{L}$ and $\mathcal{U}$ are blockwise missing and the missing patterns (including the number of missing blocks) are shared by $\mathcal{L}$ and $\mathcal{U}$, we compute a revised version of their method: the block-missingness of labelled data results in K groups of samples, and for imputing each of the missing X_{ij} in group r by $\mathbb{E}(X_{ij}|X_{iJ(r,k)})$, we fit a linear regression model for X_{ij} and the random vector $X_{iJ(r,k)}$ based on the samples in group k and the unlabelled data $\mathcal{U}$, where $J(r,k)$ is an index set of covariates that are observed in both group r and group k.

Recall that $S_0 = \{j : \beta^\star_j \neq 0\}$, and $S_0^c = \{j : \beta^\star_j = 0\}$. Define $G_1 = \{1,2,3\}, G_2 = \{4,5,6\}$. According to the settings in (**E1**) – (**E3**), G_1 is an active subset and G_2 is an inactive subset in (**E1**) – (**E3**). Tables 1 presents the ACP and AL over the sets S_0 and S_0^c, as well as the testing results for G_1 and G_2 under setting (**E1**). Tables 2–3 report the ACP and AL over the sets S_0 and S_0^c, and the testing results for G_1, G_2 under setting (**E2**). Table 4 shows the ACP and AL over the sets S_0, S_0^c, and the testing results for G_1, G_2 under setting (**E3**). The main observations are summarized as follows:

- Our pointwise estimator is generally consistent for $\beta^\star$ as the ACP and AL are reasonable across three settings, and the AL gets smaller as the labelled sample size increases. These observations are in line with the theory.

- Our proposed test has better performance than other competitors in the three examples in the sense that, our test has larger Power for testing $H_0 : \beta^\star_{G_1} = 0$, and the empirical sizes of our test for testing $H_0 : \beta^\star_{G_2} = 0$ are close to the nominal significance level 0.05.

- Table 3 indicates that our proposed method is computationally efficient and performs reasonably well when no complete labelled samples are available.

- Table 4 shows that with relatively small size of complete observations, our proposed method performs better than the two popular debiased methods, whereas they perform well when the size of complete labelled samples is large.

(n,ϵ)	Method	ACP		AL		Power	Size
		S_0	S_0^c	S_0	S_0^c	G_1	G_2
		$N=5000$					
200,$\epsilon \sim N(0,1)$	Proposed	0.877	0.954	0.897	0.880	1.000	0.040
300,$\epsilon \sim N(0,1)$	Proposed	0.907	0.952	0.728	0.716	1.000	0.050
200,$\epsilon \sim t(3)$	Proposed	0.890	0.955	1.064	1.038	1.000	0.070
300,$\epsilon \sim t(3)$	Proposed	0.905	0.953	0.906	0.888	1.000	0.080
200,$\epsilon \sim WB(0.5,0.3)$	Proposed	0.867	0.955	0.963	0.942	1.000	0.050
300,$\epsilon \sim WB(0.5,0.3)$	Proposed	0.893	0.954	0.786	0.775	1.000	0.070
		$N=1000$					
200,$\epsilon \sim N(0,1)$	Proposed	0.858	0.943	0.878	0.856	1.000	0.080
300,$\epsilon \sim N(0,1)$	Proposed	0.872	0.941	0.727	0.710	1.000	0.090
200,$\epsilon \sim t(3)$	Proposed	0.865	0.940	1.090	1.059	1.000	0.100
300,$\epsilon \sim t(3)$	Proposed	0.885	0.942	0.884	0.874	1.000	0.100
200,$\epsilon \sim WB(0.5,0.3)$	Proposed	0.860	0.942	0.927	0.906	1.000	0.050
300,$\epsilon \sim WB(0.5,0.3)$	Proposed	0.875	0.940	0.796	0.773	1.000	0.060

Table 1: Estimation results for S_0, S_0^c and hypothesis testing results for G_1, G_2 under (**E1**) with $p=450$ and varying (n,N) based on $R=100$ replications. Notes: $S_0=\{j:\beta_j^\star \neq 0\}$; $S_0^c=\{j:\beta_j^\star=0\}$; $G_1=\{1,2,3\}$; $G_2=\{4,5,6\}$.

4.2 Real Data Example

In this subsection, we apply our method to analyze the ADNI data (Mueller et al., 2005). A major goal of this study is to identify biomarkers associated with the Alzheimer's Disease (AD). This data set contains multiple measurements, such as magnetic resonance imaging (MRI), positron emission tomography (PET) imaging and other cognitive tests to track the progression of the complex disease. Similar to other relevant works (Chapman et al., 2016; Yu et al., 2020; Xue and Qu, 2020; Xue et al., 2021), we treat the mini-mental state examination (MMSE) score from cognitive tests as the response variable and use the region of interest (ROI) level data from complementary MRI and PET as the predictors. In our analysis, the variable from the MRI are subcortical volume, average cortical thickness, standard deviation of the cortical thickness and surface areas of different ROIs. These MRI variables are extracted from the MRI conducted by the Center for Imaging of Neurodegenerative Diseases at University of California, San Francisco. We further normalize the MRI variables by dividing subcortical volume, surface areas and cortical thicknesses by the sum of the ROIs' subcortical volume, sum of the ROIs' surface area and the mean cortical thickness of each participant, respectively (Xue and Qu, 2020). Other predictors from the PETs

(n, ϵ)	Method	ACP		AL		Power	Size
		S_0	S_0^c	S_0	S_0^c	G_1	G_2
$300, \epsilon \sim N(0,1)$	Proposed	0.888	0.950	0.623	0.621	1.000	0.070
	X. et al.	0.801	0.816	0.605	0.608	1.000	0.160
$600, \epsilon \sim N(0,1)$	Proposed	0.923	0.949	0.449	0.445	1.000	0.030
	X. et al.	0.894	0.895	0.457	0.454	1.000	0.070
$300, \epsilon \sim t(3)$	Proposed	0.903	0.954	0.756	0.753	1.000	0.050
	X. et al.	0.803	0.820	0.732	0.736	1.000	0.220
$600, \epsilon \sim t(3)$	Proposed	0.920	0.949	0.540	0.535	1.000	0.020
	X. et al.	0.877	0.893	0.546	0.543	1.000	0.100
$300, \epsilon \sim WB(0.5, 0.3)$	Proposed	0.878	0.950	0.668	0.665	1.000	0.040
	X. et al.	0.818	0.811	0.675	0.678	1.000	0.260
$600, \epsilon \sim WB(0.5, 0.3)$	Proposed	0.934	0.948	0.489	0.483	1.000	0.070
	X. et al.	0.906	0.897	0.510	0.508	1.000	0.100

Table 2: Estimation results for S_0, S_0^c and hypothesis testing results for G_1, G_2 under (**E2**) (i) with $p = 150, N = 5000$ and varying n based on $R = 100$ replications. Notes: $S_0 = \{j : \beta_j^\star \neq 0\}$; $S_0^c = \{j : \beta_j^\star = 0\}$; $G_1 = \{1, 2, 3\}$; $G_2 = \{4, 5, 6\}$. X. et al., a revised version of the imputation-based method developed by Xue et al. (2021). To save the computational time, we take the selected tuning parameter by 10-folds cross-validation in the first replication as the unique tuning parameter value for all 100 replications for the method "X. et al.".

include region volumes and standard uptake value ratios (SUVRs) of various ROIs, which are segmented from the PETs by the Jagust Lab at University of California, Berkeley. We also normalize region volumes by dividing them by the sum of ROI volume of each subject.

In our analysis, the observations in the third phase of the ADNI study (ADNI-3) at year 2 visit are regarded as labelled data, and the observations in ADNI-2 at year 2 visit are treated as unlabelled data. To ensure independence of the labelled and unlabelled data, the subjects in the labelled data set are removed from the unlabelled data set on the basis of the "visit code" provided by the ADNI study. We normalize the response MMSE before analysis. Overall, 172 features are from MRI and 208 from PET. There are 334 labelled subjects, which include (1) 116 participants with complete MRI and PET features; (2) 102 participants with only MRI features; (3) 116 participants with only PET features. Thus, block missingness occurs when we integrate the data for a combined analysis. Meanwhile, 334 unlabelled participants are available. Thus, $K = 3$, $n = 334$, $p = 380$, $N = 333$, $p_1 = 172$, $p_2 = 208$, $p_3 = 380$, $n_1 = 102$ and $n_2 = n_3 = 116$.

ϵ	Method	ACP		AL		Power	Size	Time
		S_0	S_0^c	S_0	S_0^c	G_1	G_2	
$\epsilon \sim N(0,1)$	Proposed	0.891	0.948	0.552	0.547	1.000	0.040	1.265
	X. et al.	0.780	0.814	0.531	0.531	1.000	0.180	37.522
$\epsilon \sim t(3)$	Proposed	0.900	0.947	0.686	0.678	1.000	0.060	1.282
	X. et al.	0.811	0.813	0.658	0.658	1.000	0.220	38.187
$\epsilon \sim WB(0.5, 0.3)$	Proposed	0.909	0.951	0.677	0.666	1.000	0.020	1.266
	X. et al.	0.827	0.819	0.679	0.679	1.000	0.140	38.281

Table 3: Estimation results for S_0, S_0^c and hypothesis testing results for G_1, G_2 under (**E2**) (ii) with $n = 300, p = 150, N = 5000$ based on $R = 50$ replications. Notes: $S_0 = \{j : \beta_j^\star \neq 0\}$; $S_0^c = \{j : \beta_j^\star = 0\}$; $G_1 = \{1, 2, 3\}$; $G_2 = \{4, 5, 6\}$. Time, in minutes. X. et al., a revised version of the imputation-based method developed by Xue et al. (2021). The tuning parameter for the method “X. et al.” was selected by 10-folds cross-validation in each replication.

To identify the important biomarkers associated with MMSE, we apply the proposed method to the resulting data set. For comparison, we also compute a revised version of the method by van de Geer et al. (2014), that uses the unlabelled data to construct the approximate inverse estimator of $\mathbf{\Sigma} = \mathbb{E}(XX^\top)$ but does not use the labelled observations with missing covariates, as well as the method by Zhang and Zhang (2014) using complete observations. The significance level $\alpha = 0.01$. The results of the identified biomarkers are reported in Table 5. Our method identifies 24 important biomarkers, which includes 8 biomarkers from MRI and 16 biomarkers from PET. Three biomarkers, namely “ST24TA”, “CTX_RH_TRANSVERSETEMPORAL_SUVR” and “CTX_LH_CUNEUS_SUVR” are identified by all methods, and 17 biomarkers are identified by our method only, e.g., “RIGHT_HIPPOCAMPUS_VOLUME”. The three biomarkers identified by all methods represent the average thickness of the left entorhinal cortex, SUVR of the right transverse temporal and SUVR of the left cuneus, respectively. Indeed, the entorhinal cortex, transverse temporal and left cuneus are related to AD (Wenk, 2003; Wang et al., 2006; Paola et al., 2007; Peters et al., 2009). The estimates of coefficients corresponding to “ST24TA”, “CTX_RH_TRANSVERSETEMPORAL_SUVR” are positive across all methods, suggesting that the two biomarkers have a positive impact on AD; in constrast, the estimate of the coefficient corresponding to “CTX_LH_CUNEUS_SUVR” is negative for all methods, which means that this biomarker negatively affects AD. However, in terms of the 95% confidence interval, the lengths of our estimation for “CTX_RH_TRANSVERSETEMPORAL_SUVR” and “CTX_LH_CUNEUS_SUVR” are smaller than the revised semi-supervised method based on van de Geer et al. (2014) and the method in Zhang and Zhang (2014) but our estimation for “ST24TA” has a bigger length than others. The biomarkers, “ST12SV”,“ST71SV”, “LEFT_AMYGD ALA_VOLUME” and

n_4	Method	ACP		AL		Power	Size
		S_0	S_0^c	S_0	S_0^c	G_1	G_2
$n_4 = 60$	Proposed	0.884	0.949	0.528	0.524	1.000	0.050
	X. et al.	0.826	0.804	0.531	0.508	1.000	0.220
	G. et al.	0.772	0.948	0.521	0.519	1.000	0.100
	Z. & Z.	0.938	0.970	1.209	1.206	0.970	0.040
$n_4 = 100$	Proposed	0.882	0.945	0.483	0.482	1.000	0.070
	X. et al.	0.831	0.834	0.490	0.473	1.000	0.130
	G. et al.	0.867	0.950	0.410	0.410	1.000	0.020
	Z. & Z.	0.936	0.945	0.578	0.579	1.000	0.080
$n_4 = 200$	Proposed	0.874	0.947	0.411	0.415	1.000	0.060
	X. et al.	0.898	0.889	0.421	0.412	1.000	0.060
	G. et al.	0.908	0.955	0.298	0.299	1.000	0.040
	Z. & Z.	0.922	0.941	0.353	0.356	1.000	0.080

Table 4: Estimation results for S_0, S_0^c and hypothesis testing results for G_1, G_2 under (**E3**) with $p = 150, N = 5000$ and varying n_4 based on $R = 100$ replications. Notes: $S_0 = \{j : \beta_j^\star \neq 0\}$; $S_0^c = \{j : \beta_j^\star = 0\}$; $G_1 = \{1, 2, 3\}$; $G_2 = \{4, 5, 6\}$; n_4, the size of complete labelled observations. X. et al., a revised version of the method developed by Xue et al. (2021); G. et al., the method by van de Geer et al. (2014); Z. & Z., the method by Zhang and Zhang (2014). To save the computational time, we take the selected tuning parameter by 10-folds cross-validation in the first replication as the unique tuning parameter value for all 100 replications for the method "X. et al.".

"RIGHT_AMYGDALA_VOLUME", related to the amygdala region, are only identified by our proposed method (Coupé et al., 2019).

In addition, hypothesis testing is conducted to identify important ROIs. The results are summarized in Table 6. Specifically, 5 biomarkers are related to the hippocampus cortex in the brain, that is, $G_1 = \{$"RIGHT_HIPPOCA-MPUS_SUVR", "RIGHT_HIPPOCAMPUS_VOLUME", "ST88SV", "LEFT_HIPPOCAMPUS_SUVR", "ST29SV"$\}$. For testing $H_0 : \beta_{G_1}^\star = 0$, the p-value of our test statistic T defined in (12) is 6.680×10^{-5}, so we reject the null hypothesis H_0, implying significant impact of the hippocampus cortex on AD. In contrast, the p-values of the test stastics in van de Geer et al. (2014) and Zhang and Zhang (2014) are around 0.030 and 0.015, respectively. We also test the effect of the entorhinal cortex on AD. The related biomarkers are $G_2 = \{$"CTX_LH_ENTORHINAL_SUVR", "CTX_LH_ENTORHINAL_VOLUME", "ST24TA", "ST24TS", "CTX_RH_ENTORHINAL_SUVR", "CTX_RH_ENTORHINAL_VOLUME", "ST83TA", "ST 83TS"$\}$. For testing $H_0 : \beta_{G_2}^\star = 0$, the p-values of all methods are very close to zero, indicating that the entorhinal cortex is also an important ROI for tracking the progression of AD.

Method	Biomarker	Est.	CI
G. et al.	ST24TA	1.140	(0.643,1.636)
	CTX_RH_TRANSVERSETEMPORAL_SUVR	1.074	(0.324,1.824)
	CTX_LH_CUNEUS_SUVR	-1.191	(-2.054,-0.328)
Z. & Z.	ST24TA	1.130	(0.521,1.739)
	CTX_RH_TRANSVERSETEMPORAL_SUVR	1.322	(0.451,2.194)
	CTX_LH_CUNEUS_SUVR	-1.684	(-2.670,-0.697)
Proposed	ST24TA	1.584	(0.822,2.347)
	CTX_RH_TRANSVERSETEMPORAL_SUVR	1.110	(0.416,1.804)
	CTX_LH_CUNEUS_SUVR	-1.291	(-2.108,-0.473)
	ST111TA	0.483	(0.095,0.872)
	ST12SV	1.327	(0.404,2.250)
	ST13TS	-0.395	(-0.676,-0.114)
	ST14TA	-0.614	(-1.003,-0.224)
	ST60TS	-0.375	(-0.652,-0.099)
	ST69SV	-0.542	(-0.820,-0.264)
	ST71SV	1.071	(0.201,1.941)
	RIGHT_PALLIDUM_SUVR	0.841	(0.232,1.450)
	CTX_RH_PARAHIPPOCAMPAL_VOLUME	-0.515	(-0.874,-0.156)
	CTX_RH_SUPERIORPARIETAL_SUVR	0.741	(0.142,1.341)
	RIGHT_THALAMUS_PROPER_VOLUME	-0.830	(-1.474,-0.186)
	LEFT_AMYGDALA_VOLUME	-0.974	(-1.724,-0.224)
	LEFT_CAUDATE_SUVR	0.932	(0.241,1.624)
	CC_MID_ANTERIOR_SUVR	-0.825	(-1.417,-0.233)
	CC_MID_POSTERIOR_SUVR	-0.623	(-1.147,-0.098)
	RIGHT_AMYGDALA_VOLUME	-0.928	(-1.496,-0.359)
	RIGHT_HIPPOCAMPUS_VOLUME	-0.739	(-1.260,-0.218)

Table 5: Estimates (Est.) and 95% confidence intervals (CI) for the coefficients of important biomarkers. Notes: G. et al., a revised version of the method by van de Geer et al. (2014); Z. & Z., the method by Zhang and Zhang (2014).

5. Concluding Remarks

In this paper, we study a semi-supervised inference for single or low-dimensional regression coefficients in a high-dimensional linear model with block-missing data, which covers the construction of simultaneous confidence intervals and hypothesis testing. Our main idea is inspired by the association between the target parameter $\beta^\star$ and the partially observed covariates, thereby eschewing any imputation. Then, the KKT conditions corresponding to (5) are utilised to develop a bias-corrected estimator with a tractable limiting distribution. In particular, it is required that the sample size of unlabelled data satisfies

Set	Method	Value of test statistic	p-value
G_1	Proposed	28.785	6.680×10^{-5}
	G. et al.	13.925	3.048×10^{-2}
	Z. & Z.	15.837	1.466×10^{-2}
G_2	Proposed	30.415	1.785×10^{-4}
	G. et al.	34.020	4.028×10^{-5}
	Z. & Z.	25.779	1.146×10^{-3}

Table 6: Hypothesis testing results for G_1, G_2 under ADNI data. Notes: G. et al., a revised version of the method by van de Geer et al. (2014); Z. & Z., the method by Zhang and Zhang (2014).

$N \gtrsim \max_k[n_k^2/\{s^{(k)}\log p_k\}]$ to ensure the limiting theory of the proposed estimator. However, the construction of $\bar{S}$ in Section 2.2 is a weighted average of SS_k, $k = 1, \ldots, K$, which may not be the optimal way of integration in terms of mean squared errors. It would be interesting to consider this problem in the future.

Acknowledgments

The authors wish to thank the action editor Professor Ji Zhu and three reviewers for their insightful comments and constructive suggestions that significantly helped improve the article. This work was conducted when S. Song was a postdoctoral fellow in Department of Statistics, The Chinese University of Hong Kong. Y. Lin's work was supported by the Hong Kong Research Grants Council (Grant No. 14306219, 14306620, 14304523), and Direct Grants for Research, The Chinese University of Hong Kong. Y. Zhou's work was supported by the State Key Program of National Natural Science Foundation of China (Grant No. 71931004) and National Natural Science Foundation of China (Grant No. 92046005) and the National Key R&D Program of China (Grant No. 2021YFA1000100, 2021YFA1000101).

Appendix A. Preliminary

We first state four lemmas that will be used later. For brevity, we suppose that Conditions (A1)-(A4) hold throughout Appendix A.

Let Z be certain subvector of X. For i.i.d. copies of Z, $\{Z_1, Z_2, \ldots, Z_m\}$, define

$$\mathcal{G}_1(e_i, e_j, t) = \left\{ \left| e_i^\top \left(\frac{1}{m} \sum_{i=1}^m Z_i Z_i^\top \right) e_j - e_i^\top \mathbb{E}(ZZ^\top) e_j \right| \lesssim \frac{t}{\sqrt{m}} \right\},$$

where $e_j \in \mathbb{R}^{p_z}$ is the j-th unit column vector and p_z represents the dimension of Z. Lemma 6 provides a general result for the subvector Z.

Lemma 6 *Suppose that Condition (B1) holds. Then, for any $t \lesssim \sqrt{m}$,*

$$\mathbb{P}\left\{\mathcal{G}_1(e_i, e_j, t)\right\} \geq 1 - 2\exp(-Ct^2),$$

where $C > 0$ is an absolute constant.

Proof Let $\|\cdot\|_{\varphi_1}$ and $\|\cdot\|_{\varphi_2}$ be the sub-exponential norm and sub-Gaussian norm of a random variable, respectively. As a subvector of X, Z is also a sub-Gaussian random vector. Thus, $e_i^\top \left\{Z_i Z_i^\top - \mathbb{E}(ZZ^\top)\right\} e_j$ is a centred random variable with sub-exponential norm

$$\left\| e_i^\top \left\{ Z_i Z_i^\top - \mathbb{E}(ZZ^\top) \right\} e_j \right\|_{\varphi_1} \leq 4\|Z_i\|_{\varphi_2}^2,$$

by Lemma 10 in Cai and Guo (2020). Let $C_1 = 4\|Z_i\|_{\varphi_2}^2$. It follows from Corollary 2.8.3 in Vershynin (2018) that there exists some constant C such that, for any $t \lesssim \sqrt{m}$,

$$\mathbb{P}\left\{ \left| e_i^\top \left(\frac{1}{m} \sum_{i=1}^m Z_i Z_i^\top \right) e_j - e_i^\top \mathbb{E}(ZZ^\top) e_j \right| \geq \frac{C_1 t}{\sqrt{m}} \right\} \leq 2\exp(-Ct^2).$$

We complete the proof. ■

Recall that for any reduced model $Y = \{Z^{(k)}\}^\top \theta^{(k)} + \delta^{(k)}$, $k = 1, \ldots, K$, $Z^{(k)}$ is a subvector of X, $\theta^{(k)}$ is the ℓ_2 projection coefficient vector of Y onto $Z^{(k)}$ and $\delta^{(k)}$ is the corresponding residual term. For i.i.d. copies of $(Z^{(k)}, \delta^{(k)})$, $\{(Z_1^{(k)}, \delta_1^{(k)}), (Z_2^{(k)}, \delta_2^{(k)}), \ldots, (Z_{n_k}^{(k)}, \delta_{n_k}^{(k)})\}$, define

$$\mathcal{G}_2^{(k)}(e_j, t) = \left\{ \left| \frac{1}{n_k} \sum_{i=1}^{n_k} e_j^\top Z_i^{(k)} \delta_i^{(k)} \right| \lesssim \frac{t}{\sqrt{n_k}} \right\}, \; \mathcal{G}_3^{(k)} = \left\{ \max_j \left| \frac{1}{n_k} \sum_{i=1}^{n_k} e_j^\top Z_i^{(k)} \delta_i^{(k)} \right| \lesssim \sqrt{\frac{\log p_k}{n_k}} \right\},$$

where p_k is the dimension of $Z^{(k)}$.

Lemma 7 *Suppose that Conditions (B1) and (C2) hold. Then, for all $k \in \{1, \ldots, K\}$, $\mathbb{P}\{\mathcal{G}_2^{(k)}(e_j, t)\} \geq 1 - 2\exp(-Ct^2)$ holds for any $t \lesssim \sqrt{n_k}$ and $\mathbb{P}\{\mathcal{G}_3^{(k)}\} \geq 1 - 2p_k \exp(-C \log p_k)$, where $C > 0$ is an absolute constant.*

Proof By Lemma 2.7.7 in Vershynin (2018) and Conditions (B1) and (C2), $e_j^\top Z_i^{(k)}\delta_i^{(k)}$ is a centred sub-exponential random variable with sub-exponential norm

$$\|e_j^\top Z_i^{(k)}\delta_i^{(k)}\|_{\varphi_1} \leq 2\|Z_i^{(k)}\|_{\varphi_2}\|\delta_i^{(k)}\|_{\varphi_2} \leq C_1\eta^{(k)},$$

where $C_1 = 2\|Z_i^{(k)}\|_{\varphi_2}\|\delta_i^{(k)}/\eta^{(k)}\|_{\varphi_2}$ and $\{\eta^{(k)}\}^2$ is the population variance of $\delta^{(k)}$. Corollary 2.8.3 in Vershynin (2018) states that for any $t \lesssim \sqrt{n_k}$,

$$\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{i=1}^{n_k} e_j^\top Z_i^{(k)}\delta_i^{(k)}\right| \geq \frac{t}{\sqrt{n_k}}\cdot C_1\eta^{(k)}\right) \leq 2\exp(-Ct^2),$$

where $C > 0$ is an absolute constant. That is, $\mathbb{P}\{\mathcal{G}_2^{(k)}(e_j, t)\} \geq 1 - 2\exp(-Ct^2)$. Observe that the event $\mathcal{G}_3^{(k)}$ holds on the event $\cap_{j=1}^{p_k}\mathcal{G}_2^{(k)}(e_j, \sqrt{\log p_k})$. Thus, $\mathbb{P}\{\mathcal{G}_3^{(k)}\} \geq 1 - 2p_k\exp(-C\log p_k)$. ■

Recall that p_k and n_k are the dimension and sample size of $Z^{(k)}$ respectively. The tuning parameter λ_k is defined in Section 2.2.

Lemma 8 *Assume that Conditions (B1)-(B2), (C2)-(C3) hold and* $\log p_k \lesssim n_k$. *Then, for* $\lambda_k \asymp \sqrt{\log p_k/n_k}$, *for* $k = 1, \dots, K$, *the events*

$$\|\hat\theta_k - \theta^{(k)}\|_1 \lesssim s^{(k)}\sqrt{\frac{\log p_k}{n_k}}, \quad \|\mathbf{Z}^{(k)}\{\hat\theta_k - \theta^{(k)}\}\|_2^2/n_k \lesssim s^{(k)}\frac{\log p_k}{n_k},$$

hold with probability at least $1 - 4p_k^{-C_k}$ *for some absolute constant* $C_k > 0$, *where the* i*-th row of* $\mathbf{Z}^{(k)}$ *is* $Z_i^{(k)}, i = 1, \dots, n_k$, *and* $s^{(k)} = \|\theta^{(k)}\|_0$.

Proof In view of Lemma 7, in the reduced model $Y_i = \{Z_i^{(k)}\}^\top\theta^{(k)} + \delta_i^{(k)}$ for $i \in \mathcal{S}(k)$ and $k = 1, \dots, K$, the subvector $Z_i^{(k)}$ of X_i is also sub-Gaussian under Condition (B1). Recall that $\mathcal{I}(k)$ is the index set of the subvector $Z^{(k)}$. It follows from Conditions (B1), (C2) and Lemma 7 that the event

$$\max_j\left|\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} e_j^\top Z_i^{(k)}\delta_i^{(k)}\right| \lesssim \sqrt{\frac{\log p_k}{n_k}},$$

holds with probability at least $1 - 2p_k^{-C_k'}$ for some absolute constant $C_k' > 0$ if $\log p_k \lesssim n_k$, where e_j is j-th unit column vector of length $|\mathcal{I}(k)|$.

We shall then show that for the nonzero index set $\mathcal{S}^{(k)}$ of $\theta^{(k)}$, $n_k^{-1}\sum_{i\in\mathcal{S}(k)} Z_i^{(k)}\{Z_i^{(k)}\}^\top$-compatibility condition holds with a large probability. To see this, observe that the event

$$\left\|\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} Z_i^{(k)}\{Z_i^{(k)}\}^\top - \mathbb{E}\left[Z^{(k)}\{Z^{(k)}\}^\top\right]\right\|_\infty \lesssim \sqrt{\frac{\log p_k}{n_k}}, \tag{13}$$

holds on the event $\cap_{j=1}^{p_k}\mathcal{G}_1(e_j, e_j, \sqrt{\log p_k})$ with $Z_i = Z_i^{(k)}$ for $i = 1, \ldots, m$ and $m = n_k$. Thus applying Lemma 6 and Condition (B1), one can easily check that the event in (13) holds with probability at least $1 - 2p_k^{-C_k''}$ for some absolute constant $C_k'' > 0$ when $\log p_k \lesssim n_k$. Moreover, the smallest eigenvalue of $\mathbb{E}\left[Z^{(k)}\{Z^{(k)}\}^\top\right]$ is bounded away from zero by Condition (B2), suggesting that its compatibility condition holds for $\mathcal{S}^{(k)}$. Hence, by Condition (C3) and Corollary 6.8 in Bühlmann and van de Geer (2011), we have shown that $n_k^{-1}\sum_{i\in\mathcal{S}(k)} Z_i^{(k)}\{Z_i^{(k)}\}^\top$-compatibility condition holds for $\mathcal{S}^{(k)}$ with probability at least $1 - 2p_k^{-C_k''}$.

Finally, Theorem 6.1 in Bühlmann and van de Geer (2011) shows that for $\lambda_k \asymp \sqrt{\log p_k / n_k}$,

$$\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\left[\{Z_i^{(k)}\}^\top\{\hat{\theta}_k - \theta^{(k)}\}\right]^2 + \lambda_k\|\hat{\theta}_k - \theta^{(k)}\|_1 \lesssim \lambda_k^2 s^{(k)},$$

holds with probability at least $1 - 4p_k^{-C_k}$ for some absolute constant $C_k = \min\{C_k', C_k''\}$. Therefore, the statement of the lemma follows. ■

For ease of presentation, we rewrite $Z_i^{(k)}$ as $X_{i\mathcal{I}(k)}$ for $i \in \mathcal{S}(k)$ and $k = 1, \ldots, K$. Now, $\widetilde{\mathbf{\Sigma}}_{n,N}^{(k)}$ can be written as

$$\frac{1}{N + n_k}\left\{\sum_{i=n+1}^{n+N} X_{i\mathcal{I}(k)}X_{i\mathcal{I}(k)}^\top + \sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}X_{i\mathcal{I}(k)}^\top\right\}.$$

Write $\mathbf{\Sigma}^{(k)} = \mathbb{E}\{X_{i\mathcal{I}(k)}X_{i\mathcal{I}(k)}^\top\}$ and $\widehat{\mathbf{\Sigma}}^{(k)} = n_k^{-1}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}X_{i\mathcal{I}(k)}^\top$.

Lemma 9 *Suppose that Conditions (B1)-(B3) and (C2)-(C3) hold and* $\log p_k \lesssim \min\{N, n_k\}$. *If* $\lambda_k \asymp \sqrt{\log p_k / n_k}$ *for* $k = 1, 2, \ldots, K$, *then,*

$$SS_k - \mathbf{\Sigma}^{(k)}\theta^{(k)} = \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}\delta_i^{(k)} + \Delta_k,$$

where

$$\|\Delta_k\|_\infty \lesssim \frac{N}{(N + n_k)}\left(\sqrt{\frac{\log p_k}{n_k}} + \sqrt{\frac{\log p_k}{N}}\right) s^{(k)}\sqrt{\frac{\log p_k}{n_k}} + \sqrt{\frac{s^{(k)}\log p_k}{(N + n_k)}}, \tag{14}$$

holds with probability at least $1 - 8p_k^{-C_k}$ *for some absolute constant* $C_k > 0$.

Proof Observe that

$$
\begin{aligned}
& SS_k - \mathbf{\Sigma}^{(k)}\theta^{(k)} \\
&= \widetilde{\mathbf{\Sigma}}^{(k)}_{n,N}\hat{\theta}^{(k)} - \mathbf{\Sigma}^{(k)}\theta^{(k)} + \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}\{Y_i - X^{\top}_{i\mathcal{I}(k)}\hat{\theta}^{(k)}\} \\
&= \widetilde{\mathbf{\Sigma}}^{(k)}_{n,N}\{\hat{\theta}^{(k)} - \theta^{(k)}\} + \{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\}\theta^{(k)} + \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}\{X^{\top}_{i\mathcal{I}(k)}\theta^{(k)} + \delta_i^{(k)} - X^{\top}_{i\mathcal{I}(k)}\hat{\theta}^{(k)}\} \\
&= \{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)}\}\{\hat{\theta}^{(k)} - \theta^{(k)}\} + \{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\}\theta^{(k)} + \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}\delta_i^{(k)}.
\end{aligned}
$$

It suffices to derive the convergence rate of $\|\{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)}\}\{\hat{\theta}^{(k)} - \theta^{(k)}\}\|_\infty$ and $\|\{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\}\theta^{(k)}\|_\infty$, respectively. Then, in view of the fact that

$$
\begin{aligned}
\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)} &= \frac{1}{N+n_k}\sum_{i=n+1}^{n+N} X_{i\mathcal{I}(k)}X^{\top}_{i\mathcal{I}(k)} + \left(\frac{1}{N+n_k} - \frac{1}{n_k}\right)\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}X^{\top}_{i\mathcal{I}(k)} \\
&= \frac{N}{N+n_k}\left\{\frac{1}{N}\sum_{i=n+1}^{n+N} X_{i\mathcal{I}(k)}X^{\top}_{i\mathcal{I}(k)} - \mathbf{\Sigma}^{(k)} + \mathbf{\Sigma}^{(k)} - \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{i\mathcal{I}(k)}X^{\top}_{i\mathcal{I}(k)}\right\},
\end{aligned}
$$

it follows from Lemma 6 and Condition (B1) that the event

$$
\|\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)}\|_\infty \lesssim \frac{N}{(N+n_k)}\left(\sqrt{\frac{\log p_k}{n_k}} + \sqrt{\frac{\log p_k}{N}}\right),
$$

holds with probability at least $1 - 2p_k^{-C'_k}$ for some absolute constant $C'_k > 0$ when $\log p_k \lesssim \min\{N, n_k\}$. Moreover, it has been shown in Lemma 8 that, under Conditions (B1)-(B2) and (C2)-(C3), $\|\hat{\theta}^{(k)} - \theta^{(k)}\|_1 \lesssim s^{(k)}\sqrt{\log p_k/n_k}$ holds with probability at least $1 - 4p_k^{-C''_k}$ for some absolute constant $C''_k > 0$. It then follows that

$$
\begin{aligned}
\|\{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)}\}\{\hat{\theta}^{(k)} - \theta^{(k)}\}\|_\infty &\le \|\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \widehat{\mathbf{\Sigma}}^{(k)}\|_\infty\|\hat{\theta}^{(k)} - \theta^{(k)}\|_1 \\
&\lesssim \frac{N}{(N+n_k)}\left(\sqrt{\frac{\log p_k}{n_k}} + \sqrt{\frac{\log p_k}{N}}\right)s^{(k)}\sqrt{\frac{\log p_k}{n_k}},
\end{aligned}
$$

holds with probability at least $1 - 6p_k^{-C_k}$ for $C_k = \min\{C'_k, C''_k\}$. On the other hand, Lemma 6 and Condition (B1) imply that $\|\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\|_\infty \lesssim \sqrt{\log p_k/(n_k+N)}$ holds with probability at least $1 - 2p_k^{-C'_k}$, and it is easy to check that

$$
\|\theta^{(k)}\|_1 \le \sqrt{s^{(k)}}\|\theta^{(k)}\|_2 \le \sqrt{s^{(k)}\mathbb{E}(Y^2)}/\min\{\sqrt{\mathbb{E}(Y^2)}, \Lambda_{min}\},
$$

suggesting that under Conditions (B1) and (B3),

$$
\begin{aligned}
\|\{\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\}\theta^{(k)}\|_\infty &\le \|\widetilde{\mathbf{\Sigma}}^{(k)}_{n,N} - \mathbf{\Sigma}^{(k)}\|_\infty\|\theta^{(k)}\|_1 \\
&\lesssim \sqrt{\frac{s^{(k)}\log p_k}{(N+n_k)}},
\end{aligned}
$$

holds with probability at least $1 - 2p_k^{-C_k'}$. Hence, we have proved that (14) holds with probability at least $1 - 8p_k^{-C_k}$. The proof is complete. ■

Appendix B. Proofs for Main Text

Proof of Theorem 2. Our proof is structurally similar to that of Theorem 6.1 in Bühlmann and van de Geer (2011). It follows from the definition of $\hat{\beta}_\lambda$ in (5) that the following basic inequality holds:

$$\hat{\beta}_\lambda^\top \widehat{\mathbf{\Sigma}}_N \hat{\beta}_\lambda - 2\bar{S}^\top \hat{\beta}_\lambda + \lambda \|\hat{\beta}_\lambda\|_1 \leq (\beta^\star)^\top \widehat{\mathbf{\Sigma}}_N \beta^\star - 2\bar{S}^\top \beta^\star + \lambda\|\beta^\star\|_1.$$

Then,

$$(\hat{\beta}_\lambda - \beta^\star)^\top \widehat{\mathbf{\Sigma}}_N (\hat{\beta}_\lambda - \beta^\star) + \lambda\|\hat{\beta}_\lambda\|_1 \leq 2(\bar{S} - \widehat{\mathbf{\Sigma}}_N \beta^\star)^\top (\hat{\beta}_\lambda - \beta^\star) + \lambda\|\beta^\star\|_1.$$

Recall that $\boldsymbol{J}$ is a $p \times p$ diagonal matrix with $\boldsymbol{J}_{j,j} = \sum_{k\in\mathcal{H}(j)} \sqrt{n_k}$, where $\mathcal{H}(j)$ stands for the collection of groups with the j-th element of X in $\mathcal{L}$ observed. Define $w_{k,j} = \sqrt{n_k}/\sum_{k\in\mathcal{H}(j)}\sqrt{n_k}$ for $k = 1, \ldots, K, j = 1, \ldots, p$. One can show that $\bar{S} - \widehat{\mathbf{\Sigma}}_N\beta^\star = (\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_N)\beta^\star + \boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w + \Delta_0$, where $\mathbf{X}_{fill}$ and δ_w can be visualized in Figure 3 and $\Delta_0 = (\Delta_{0,1}, \ldots, \Delta_{0,p})^\top$,

$$\Delta_{0,j} = \sum_{k\in\mathcal{H}(j)} w_{k,j}\Delta_{k,(j)}, \tag{15}$$

$\Delta_{k,(j)}$ denotes the element of Δ_k corresponding to the position of the j-th element of X in $Z^{(k)}$ if $j \in \mathcal{I}(k)$. In other words,

$$\begin{aligned}(\hat{\beta}_\lambda - \beta^\star)^\top \widehat{\mathbf{\Sigma}}_N (\hat{\beta}_\lambda - \beta^\star) + \lambda\|\hat{\beta}_\lambda\|_1 \leq 2(\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w)^\top(\hat{\beta}_\lambda - \beta^\star) + \lambda\|\beta^\star\|_1 \\ + 2\left\{(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_N)\beta^\star + \Delta_0\right\}^\top (\hat{\beta}_\lambda - \beta^\star).\end{aligned} \tag{16}$$

Let

$$\mathcal{F}_0(e_j, t) = \left\{ \left| e_j^\top \boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top \delta_w \right| \lesssim \sum_{k\in\mathcal{H}(j)} \frac{t}{\sum_{k\in\mathcal{H}(j)}\sqrt{n_k}} \right\},$$

$$\mathcal{F}_0 = \left\{ \max_{1\leq j\leq p} \left| e_j^\top \boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w \right| \lesssim \lambda_0 \right\},$$

where $\lambda_0 \asymp \max_j[|\mathcal{H}(j)|\sqrt{\log p}/\{\sum_{k\in\mathcal{H}(j)}\sqrt{n_k}\}]$ and $|\mathcal{H}(j)|$ denotes the cardinality of the set $\mathcal{H}(j)$. Theorem 2.8.2 in Vershynin (2018) tells that there exists a constant $C_1 > 0$ such that for any $t \lesssim w_{k,j}\sqrt{n_k}$, $\mathbb{P}\{\mathcal{G}_4^{(k)}(e_j, t)\} \geq 1 - 2\exp(-C_1 t^2/w_{k,j}^2)$ holds, where

$$\mathcal{G}_4^{(k)}(e_j, t) = \left\{ \left| \frac{w_{k,j}}{n_k} \sum_{i\in\mathcal{S}(k)} e_j^\top Z_i^{(k)} \delta_i^{(k)} \right| \lesssim \frac{t}{\sqrt{n_k}} \right\}.$$

Note that by the definition of $\mathbf{X}_{fill}, \delta_w$ and $w_{k,j}$, on the event $\cap_{k=1}^K \mathcal{G}_4^{(k)}(e_j, w_{k,j}t)$, the event $\mathcal{F}_0(e_j,t)$ holds and hence we have $\mathbb{P}\{\mathcal{F}_0(e_j,t)\} \geq 1-2K\exp(-C_1 t^2)$ for any $t \lesssim \min_k \sqrt{n_k}$. Thus, $\cap_{j=1}^p \mathcal{F}_0(e_j,\sqrt{\log p}) \subset \mathcal{F}_0$ implies that $\mathbb{P}(\mathcal{F}_0) \geq 1-2Kp\exp\{-C_1\ (\sqrt{\log p})^2\} \geq 1-2Kp^{-C_2}$ for some constant $C_2>0$ (this result holds if $\log p \lesssim \min_k n_k$). On $\mathcal{F}_0$ and $\lambda \geq 4\lambda_0$, by (16), we have shown that

$$
\begin{aligned}
2(\hat{\beta}_\lambda-\beta^\star)^\top \widehat{\mathbf{\Sigma}}_N(\hat{\beta}_\lambda-\beta^\star)+2\lambda\|\hat{\beta}_\lambda\|_1 \leq \lambda\|\hat{\beta}_\lambda-\beta^\star\|_1+2\lambda\|\beta^\star\|_1 \\
+4\left\{(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_N)\beta^\star+\Delta_0\right\}^\top(\hat{\beta}_\lambda-\beta^\star).
\end{aligned}
$$

Then, we can show along the same lines as those of Lemma 6.3 in Bühlmann and van de Geer (2011) that

$$
\begin{aligned}
2(\hat{\beta}_\lambda-\beta^\star)^\top \widehat{\mathbf{\Sigma}}_N(\hat{\beta}_\lambda-\beta^\star)+\lambda\|\hat{\beta}_{\lambda,S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{\lambda,S_0}-\beta_{0,S_0^c}\|_1 \\
+4\left\{(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_N)\beta^\star+\Delta_0\right\}^\top(\hat{\beta}_\lambda-\beta^\star). \qquad (17)
\end{aligned}
$$

where the j-th element of $\hat{\beta}_{\lambda,S}$ is $\hat{\beta}_{\lambda,j}I(j\in S)$ for a certain set S and $\beta_{0,S}$ is defined in the same manner.

Next, we intend to prove that the $\widehat{\mathbf{\Sigma}}_N$-compatibility condition holds for the active set $S_0=\{j:\beta_j^\star\neq 0\}$. Note that the $\mathbf{\Sigma}$-compatibility condition holds for the set S_0 under Condition (B2). Moreover, Lemma 6 and Condition (B1) together imply that $\|\widehat{\mathbf{\Sigma}}_N-\mathbf{\Sigma}\|_\infty \lesssim \sqrt{\log p/N}$ holds with probability at least $1-2p^{-C_3}$ for some absolute constant $C_3>0$ if $\log p \lesssim N$. When there exists a positive constant C_4 such that $s_0\sqrt{\log p/N} \leq C_4$ (this condition holds if $s_0 \ll \sqrt{N/\log p}$), it then follows from Corollary 6.8 in Bühlmann and van de Geer (2011) that $\widehat{\mathbf{\Sigma}}_N$-compatibility condition holds with probability at least $1-2p^{-C_3}$ with compatibility constant ϕ_0 for the set S_0.

Finally, by (17),

$$
(\hat{\beta}_\lambda-\beta^\star)^\top \widehat{\mathbf{\Sigma}}_N(\hat{\beta}_\lambda-\beta^\star)+\lambda\|\hat{\beta}_\lambda-\beta^\star\|_1 \leq 4\lambda^2 s_0/\phi_0^2+4\left\{(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_N)\beta^\star+\Delta_0\right\}^\top(\hat{\beta}_\lambda-\beta^\star).
$$

Under Conditions (B1), (B3), the same arguments as in the proof of Lemma 9 yield that

$$
\|(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_N)\beta^\star\|_\infty \leq \|\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_N\|_\infty\|\beta^\star\|_1 \lesssim \sqrt{s_0\log p/N}, \qquad (18)
$$

holds with probability at least $1-2p^{-C_3}$, since

$$
\|\beta^\star\|_1 \leq \sqrt{s^0}\|\beta^\star\|_2 \leq \sqrt{s^0\mathbb{E}(Y^2)}/\min\{\sqrt{\mathbb{E}(Y^2)},\Lambda_{min}\}. \qquad (19)
$$

Recall that

$$
\tau(p_k,s^{(k)},n_k,N)=\left(\frac{N}{(N+n_k)}\sqrt{\frac{\log p_k}{n_k}}+\frac{\sqrt{N\log p_k}}{(N+n_k)}\right)s^{(k)}\sqrt{\frac{\log p_k}{n_k}}+\sqrt{\frac{s^{(k)}\log p_k}{(N+n_k)}}.
$$

Furthermore, Lemma 9 implies that under Conditions (B1)-(B3) and (C2)-(C3),

$$
\|\Delta_k\|_\infty \lesssim \tau(p_k,s^{(k)},n_k,N),
$$

holds with probability at least $1 - 8p_k^{-C_k'}$ for some absolute constant $C_k' > 0$. Combining (18) and the definition of Δ_0, we have

$$\begin{aligned}
&\|(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\beta^\star\|_\infty + \|\Delta_0\|_\infty \\
&\lesssim \sqrt{s_0 \log p/N} + \max_j \left\{ \frac{1}{\sum_{k\in\mathcal{H}(j)} \sqrt{n_k}} \sum_{k\in\mathcal{H}(j)} \sqrt{n_k}\|\Delta_k\|_\infty \right\} \\
&\lesssim \sqrt{s_0 \log p/N} + \max_j [\{ \sum_{k\in\mathcal{H}(j)} \sqrt{n_k}\}^{-1} \sum_{k\in\mathcal{H}(j)} \{\sqrt{n_k}\tau(p_k, s^{(k)}, n_k, N)\}],
\end{aligned}$$

holds with probability at least $1 - 2p^{-C_3} - 8K\{\min_k p_k\}^{-C_5}$ for $C_5 = \min\{C_1', \ldots, C_K'\} > 0$ under Condition (C1). Therefore, if additionally, λ satisfies those condition in Theorem 2, we have

$$\begin{aligned}
&(\hat\beta_\lambda - \beta^\star)^\top \widehat{\boldsymbol{\Sigma}}_N (\hat\beta_\lambda - \beta^\star) \lesssim \lambda^2 s_0, \\
&\|\hat\beta_\lambda - \beta^\star\|_1 \lesssim \lambda s_0,
\end{aligned}$$

hold with probability larger than $1 - 2p^{-C_3} - 10K\{\min_k p_k\}^{-C_6}$ for $C_6 = \min\{C_2, C_5\}$. The proof of Theorem 2 is complete.

Proof of Theorem 3. The KKT conditions corresponding to (5) are

$$-(\bar S - \widehat{\boldsymbol{\Sigma}}_N \hat\beta_\lambda) + \lambda\hat\kappa = 0.$$

It then follows from Lemma 9 that $\bar S - \widehat{\boldsymbol{\Sigma}}_N\beta^\star = (\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\beta^\star + \boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w + \Delta_0$, where Δ_0 has been defined in (15). This implies that

$$\begin{aligned}
\hat\beta_\lambda - \beta^\star + \widehat{\boldsymbol{\Theta}}\lambda\hat\kappa &= \widehat{\boldsymbol{\Theta}}\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w + \widehat{\boldsymbol{\Theta}}\Delta_0 - (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_N - \mathbf{I})(\hat\beta_\lambda - \beta^\star) - \widehat{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma})\beta^\star \\
&= \widehat{\boldsymbol{\Theta}}\boldsymbol{J}^{-1}\mathbf{X}_{fill}^\top\delta_w + \widehat{\boldsymbol{\Theta}}\Delta_0 + \Delta_1 + \Delta_2,
\end{aligned}$$

where $\mathbf{I} \in \mathbb{R}^{p\times p}$ is an identity matrix, $\Delta_1 = -(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_N - \mathbf{I})(\hat\beta_\lambda - \beta^\star)$ and $\Delta_2 = -\widehat{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma})\beta^\star$. When the sample size in each group is of the same order, i.e., $n_1 \asymp n_2 \asymp \ldots \asymp n_K \asymp n$, it suffices to prove that $\sqrt{n}\|\widehat{\boldsymbol{\Theta}}\Delta_0\|_\infty = o_p(1)$, $\sqrt{n}\|\Delta_1\|_\infty = o_p(1)$ and $\sqrt{n}\|\Delta_2\|_\infty = o_p(1)$. The proof is carried out in several steps.

Step 1: To prove the convergence rate of $\|\widehat{\boldsymbol{\Theta}}_{j\cdot} - \boldsymbol{\Theta}_{j\cdot}\|_1$, $\|\widehat{\boldsymbol{\Theta}}_{j\cdot} - \boldsymbol{\Theta}_{j\cdot}\|_2$ *and* $|\hat\tau_j^2 - \tau_j^2|$. By the same arguments as in the proof of Theorem 2.4 in van de Geer et al. (2014), when using N observations of X, i.e., $\{X_{n+1}, X_{n+2}, \ldots, X_{n+N}\}$, to estimate $\widehat{\boldsymbol{\Theta}}$ and $\hat\tau_j$, we can show that $\|\widehat{\boldsymbol{\Theta}}_{j\cdot} - \boldsymbol{\Theta}_{j\cdot}\|_1 = \mathcal{O}_p(s_j\sqrt{\log p/N})$, $\|\widehat{\boldsymbol{\Theta}}_{j\cdot} - \boldsymbol{\Theta}_{j\cdot}\|_2 = \mathcal{O}_p(\sqrt{s_j \log p/N})$ and $|\hat\tau_j^2 - \tau_j^2| = \mathcal{O}_p(\sqrt{s_j\log p/N})$ uniformly in j for $\lambda_j^{(u)} \asymp \sqrt{\log p/N}$ under Conditions (B1)-(B2), (D1)-(D2).

Step 2: To derive the convergence rate of $\sqrt{n}\|\Delta_1\|_\infty$. Note that if $N \gtrsim \max_k n_k^2/\{s^{(k)} \log p_k\}$, we have $\tau(p_k, s^{(k)}, n_k, N) \asymp s^{(k)}\log p_k/n_k$. It then follows from Theorem 2 that for $n_1 \asymp n_2 \asymp \ldots \asymp n_K \asymp n$ and $s_0 \ll \min\{N/n, \sqrt{N/\log p}\}$, $\|\hat\beta_\lambda - \beta^\star\|_1 = \mathcal{O}_p(s_0\sqrt{\log p/n})$ under Conditions (B1)-(B3), (C1)-(C2), (D3) and N satisfies $N \gtrsim \max_k n^2/\{s^{(k)}\log p_k\}$. The KKT conditions for the nodewise lasso regression imply that

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_N - \mathbf{I}\|_\infty \le \max_j \lambda_j^{(u)}/\hat\tau_j^2.$$

Observe that

$$\begin{aligned}\sqrt{n}\|\Delta_1\|_\infty &\le \sqrt{n}\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_N - \mathbf{I}\|_\infty \cdot \|\hat{\beta}_\lambda - \beta^\star\|_1 \\ &\le \sqrt{n}\max_j \lambda_j^{(u)}/\hat{\tau}_j^2 \cdot \|\hat{\beta}_\lambda - \beta^\star\|_1 \\ &= O_p\left(\sqrt{\log p/N} s_0 \sqrt{\log p}\right).\end{aligned}$$

Henece, if $s_0 \ll \sqrt{N}/\log p$, then $\sqrt{n}\|\Delta_1\|_\infty = o_p(1)$.

Step 3: To derive the convergence rate of $\sqrt{n}\|\Delta_2\|_\infty$. Under Condition (B2), $C_1\|\beta^\star\|_2^2 \le (\beta^\star)^\top \boldsymbol{\Sigma}\beta^\star$, where the smallest eigenvalue of $\boldsymbol{\Sigma}$, $\Lambda_{min} \ge C_1$. By Conditions (B2), (B3) and (19), we have $\|\beta^\star\|_1 \lesssim \sqrt{s_0}$. By Lemma 6 and Condition (B1), $\|\widehat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}\|_\infty = \mathcal{O}_p(\sqrt{\log p/N})$, indicating that $\|(\widehat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma})\beta^\star\|_\infty \le \|\widehat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}\|_\infty\|\beta^\star\|_1 = \mathcal{O}_p(\sqrt{s_0\log p/N})$.

On the other hand, using the lasso for nodewise regression, we have constructed the approximate inverse $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Sigma}$ with $\widehat{\boldsymbol{\Theta}}_{j\cdot} = \widehat{C}_j/\hat{\tau}_j^2$, where $\widehat{C}_j = (-\hat{\gamma}_{j,1}, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, -\hat{\gamma}_{j,p})^\top$ for $j = 1, \ldots, p$. Note that $\|\widehat{C}_j\|_1 = 1 + \|\hat{\gamma}_j\|_1 \le 1 + \|\gamma_j\|_1 + \|\hat{\gamma}_j - \gamma_j\|_1$. Under Conditions (B1)-(B2) and (D1), one can show that $\|\hat{\gamma}_j - \gamma_j\|_1 = \mathcal{O}_p(s_j\lambda_j)$ by the same arguments as in the proof of Lemma 8. Moreover,

$$\|\gamma_j\|_1 \le \sqrt{s_j}\|\gamma_j\|_2 \le \sqrt{s_j\boldsymbol{\Sigma}_{j,j}/\Lambda_{min}},$$

where Λ_{min} is the smallest eigenvalue of $\boldsymbol{\Sigma}$. This suggests that $\|\widehat{C}_j\|_1 = \mathcal{O}_p(\sqrt{s_j})$ under Conditions (B2) and (D2). Moreover, we have shown in *Step 1* that $|\hat{\tau}_j^2 - \tau_j^2| = \mathcal{O}_p(\sqrt{s_j\log p/N})$ uniformly in all j; and under Condition (B2), $\tau_j^2 = 1/\boldsymbol{\Theta}_{j,j}$ is bounded away from zero. They imply that $\max_j \|\widehat{\boldsymbol{\Theta}}_{j\cdot}\|_1 = \mathcal{O}_p(\max_j \sqrt{s_j})$. It then follows that

$$\begin{aligned}\sqrt{n}\|\Delta_2\|_\infty &\le \sqrt{n}\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\beta^\star\|_\infty \cdot \max_j \|\widehat{\boldsymbol{\Theta}}_{j\cdot}\|_1 \\ &= O_P\left(\max_j \sqrt{s_0 s_j n\log p/N}\right).\end{aligned}$$

Here, if $s_0 \ll N/(\max_j s_j n\log p)$, $\sqrt{n}\|\Delta_2\|_\infty = o_p(1)$.

Step 4: To derive the convergence rate of $\sqrt{n}\|\widehat{\boldsymbol{\Theta}}\Delta_0\|_\infty$. According to the arguments in *Step 3*, $\max_j \|\widehat{\boldsymbol{\Theta}}_{j\cdot}\|_1 = \mathcal{O}_p(\max_j \sqrt{s_j})$ under Conditions (B1)-(B2) and (D1)-(D2). By Lemma 9, Conditions (B1)-(B3), (C2) and (D3), $\|\Delta_k\|_\infty = O_p\{s^{(k)}\log p_k/n_k\}$, if $N \gtrsim \max_k n_k^2/\{s^{(k)}\log p_k\}$. Note that when $n_1 \asymp n_2 \asymp \ldots \asymp n_K \asymp n$, under Condition (C1),

$$\begin{aligned}\sqrt{n}\|\widehat{\boldsymbol{\Theta}}\Delta_0\|_\infty &\le \sqrt{n}\max_j \|\widehat{\boldsymbol{\Theta}}_{j\cdot}\|_1\|\Delta_0\|_\infty \\ &\le \sqrt{n}\max_j \|\widehat{\boldsymbol{\Theta}}_{j\cdot}\|_1 \max_k \|\Delta_k\|_\infty \\ &= O_P\left[\max_j \sqrt{s_j}\max_k\{s^{(k)}\log p_k/\sqrt{n}\}\right].\end{aligned}$$

Thus, if s_j satisfies $\max_j \sqrt{s_j}\max_k\{s^{(k)}\log p_k/\sqrt{n}\} = o(1)$, $\sqrt{n}\|\widehat{\boldsymbol{\Theta}}\Delta_0\|_\infty = o_p(1)$.

As a result, under Conditions (B1)-(B3), (C1)-(C2) and (D1)-(D3), when $n_1 \asymp n_2 \asymp \ldots \asymp n_K \asymp n$, $s_0 \ll \min\{N/n, \sqrt{N}/\log p, N/(\max_j s_j n\log p)\}$ for $N \gtrsim \max_k n^2/\{s^{(k)}\log p_k\}$, and s_j satisfies $\max_j \sqrt{s_j} \max_k\{s^{(k)} \log p_k/\sqrt{n}\} = o(1)$, we have $\sqrt{n}\|\widehat{\boldsymbol{\Theta}}\Delta_0\|_\infty = o_p(1)$, $\sqrt{n}\|\Delta_1\|_\infty = o_p(1)$ and $\sqrt{n}\|\Delta_2\|_\infty = o_p(1)$. The proof of Theorem 3 is complete.

Proof of Theorem 5. Write

$$\begin{aligned}
\frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} X_{ij}\}^2 &= \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\delta_i^{(k)} + \hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 X_{ij}^2 \\
&= \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\delta_i^{(k)} X_{ij}\}^2 + \frac{2}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \delta_i^{(k)}\{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\} X_{ij}^2 \\
&\quad + \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 X_{ij}^2 \\
&=: \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\delta_i^{(k)} X_{ij}\}^2 + \Pi_1 + \Pi_2.
\end{aligned}$$

Under Condition (B1), $\{X_{ij}\}_{i \in \mathcal{S}(k)}$ are i.i.d. sub-Gaussian random variables for any $j \in \mathcal{I}(k)$. This implies that $\mathbb{E}(X_{ij}^8)$ is bounded by a constant $C_1 > 0$, where C_1 depends only on the sub-Gaussian norm of X_{ij}. Analogously, under Condition (C2), $\mathbb{E}[\{\delta_i^{(k)}\}^4]$ is bounded by a constant $C_2 > 0$ depending on the sub-Gaussian norm of $\delta_i^{(k)}$. Then, by the Cauchy–Schwartz inequality, we have

$$\begin{aligned}
\Pi_1^2 &\leq \frac{4}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 \times \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\delta_i^{(k)} X_{ij}^2\}^2 \\
&\leq \frac{4}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 \times \left[\frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\delta_i^{(k)}\}^4 \times \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} X_{ij}^8\right]^{1/2} \\
&= O_p\left(\frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2\right).
\end{aligned}$$

Moreover, under Conditions (B1)-(B2), (C2) and (D3), it follows from Lemma 8 and the definition of $\hat{\delta}_i^{(k)}$ that

$$\frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 \asymp \|\mathbf{Z}^{(k)}\{\hat{\theta}_k - \theta^{(k)}\}\|_2^2/n_k = o_p(n_k^{-1/2}), \tag{20}$$

where $\mathbf{Z}^{(k)} \in \mathbb{R}^{n_k \times p_k}$ is the design matrix with rows $\{X_{i\mathcal{I}(k)} \in \mathbb{R}^{p_k} : i \in \mathcal{S}(k)\}$, and the last equality holds under Condition (D3). Hence, $\Pi_1 = o_p(1)$. We will next show that $\Pi_2 = o_p(1)$. To see this, by Hölder's inequality, we have

$$\begin{aligned}
\Pi_2 &\leq \max_{i \in \mathcal{S}(k)} X_{ij}^2 \cdot \frac{1}{n_k - \hat{s}_k} \sum_{i \in \mathcal{S}(k)} \{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 \\
&= o_p\left\{n_k^{-1/2} \max_{i \in \mathcal{S}(k)} X_{ij}^2\right\},
\end{aligned}$$

where in the last equality holds in view of (20). By Lemma 2.2.2 in van der Vaart and Wellner (1996), we have $\|\max_{i\in\mathcal{S}(k)} X_{ij}^2\|_{\varphi_1} \leq C_3 \log n_k$, where $\|\cdot\|_{\varphi_1}$ represents the sub-exponential norm of a random variable and C_3 is a positive constant only depending on the sub-Gaussian norm of X_{ij}. This means that, for any $t > 0$,

$$\mathbb{P}\left\{\frac{1}{\log n_k}\max_{i\in\mathcal{S}(k)} X_{ij}^2 \geq t\right\} \leq 2\exp(-t/C_4),$$

where C_4 is a positive constant. This implies that $\max_{i\in\mathcal{S}(k)} X_{ij}^2 = \mathcal{O}_p(\log^2 n_k)$ under Condition (B1). Hence, $\Pi_2 = o_p(1)$. Moreover, the law of large numbers implies that

$$\frac{1}{n_k - \hat{s}_k}\sum_{i\in\mathcal{S}(k)}\{\delta_i^{(k)} X_{ij}\}^2 = \mathbb{E}\{|X_{ij}^{(k)}\delta_i^{(k)}|^2\} + o_p(1).$$

We then conclude that

$$\frac{1}{n_k - \hat{s}_k}\sum_{i\in\mathcal{S}(k)}\{\hat{\delta}_i^{(k)} X_{ij}\}^2 = \mathbb{E}\{|X_{ij}^{(k)}\delta_i^{(k)}|^2\} + o_p(1).$$

Similarly,

$$\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\hat{\delta}_i^{(k)} X_{ij} = \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\delta_i^{(k)} X_{ij} + \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}X_{ij}.$$

By Cauchy–Schwartz inequality, we have shown that

$$\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}X_{ij} \leq \left[\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}^2 \cdot \frac{1}{n_k}\sum_{i\in\mathcal{S}(k)} X_{ij}^2\right]^{1/2},$$

suggesting that

$$\frac{1}{n_k}\sum_{i\in\mathcal{S}(k)}\{\hat{\delta}_i^{(k)} - \delta_i^{(k)}\}X_{ij} = o_p(1)$$

by (20) and Condition (B1). We complete the proof.

References

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1):209–239, 2004.

Pierre C. Bellec, Arnak S. Dalalyan, Edwin Grappin, and Quentin Paris. On the prediction loss of the lasso in the partially labeled setting. *Electronic Journal of Statistics*, 12(2): 3443–3472, 2018.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758, 2019.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

Peter Bühlmann and Sara van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications.* Springer Science & Business Media, New York, 2011.

Peter Bühlmann and Sara van de Geer. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473, 2015.

T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

T. Tony Cai and Zijian Guo. Semi-supervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):391–419, 2020.

Tianxi Cai, T. Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514): 621–633, 2016.

Tianxi Cai, T Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):669–719, 2021.

Vittorio Castelli and Thomas M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

Vittorio Castelli and Thomas M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised Learning.* MIT Press, Cambridge, MA, 2009.

Kimberly R. Chapman, Hanaan Bing-Canar, Michael L. Alosco, Eric G. Steinberg, Brett Martin, Christine Chaisson, Neil Kowall, Yorghos Tripodis, and Robert A. Stern. Mini mental state examination and logical memory scores for entry into Alzheimer's disease trials. *Alzheimer's Research & Therapy*, 8(1):1–11, 2016.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Pierrick Coupé, José Vicente Manjón, Enrique Lanuza, and Gwenaelle Catheline. Lifespan changes of the human brain in alzheimer's disease. *Scientific Reports*, 9(1):1–12, 2019.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal semi-supervised estimation and inference for high-dimensional linear regression. *arXiv preprint arXiv:2011.14185*, 2020.

Jason Ernst, Qasim K. Beg, Krin A. Kay, Gábor Balázsi, Zoltán N. Oltvai, and Ziv Bar-Joseph. A semi-supervised method for predicting transcription factor–gene interactions in escherichia coli. *PLoS Computational Biology*, 4(3):e1000044, 2008.

Jessica L. Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):579–594, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Brent A. Johnson, D. Y. Lin, and Donglin Zeng. Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103(482):672–680, 2008.

Rie Johnson and Tong Zhang. Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, 54(1):275–288, 2008.

Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modeling. *Statistical Science*, 22(2):189–205, 2007.

Qi Long and Brent A. Johnson. Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics*, 16(3):596–610, 2015.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15 (4):869, 2005.

Shinichi Nakagawa and Robert P. Freckleton. Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, 23(11):592–596, 2008.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2): 103–134, 2000.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high-dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

M. D. Paola, E. Macaluso, G. A. Carlesimo, F. Tomaiuolo, K. J. Worsley, L. Fadda, and C. Caltagirone. Episodic memory impairment in patients with Alzheimer's disease is correlated with entorhinal cortex atrophy. *Journal of Neurology*, 254(6):774–781, 2007.

Frederic Peters, Fabienne Collette, Christian Degueldre, Virginie Sterpenich, Steve Majerus, and Eric Salmon. The neural correlates of verbal short-term memory in Alzheimer's disease: An fMRI study. *Brain*, 132(7):1833–1846, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics.* Springer Science & Business Media, New York, 1996.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, Cambridge, UK, 2018.

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8:1867–1891, 2007.

Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, 2008.

Liang Wang, Yufeng Zang, Yong He, Meng Liang, Xinqing Zhang, Lixia Tian, Tao Wu, Tianzi Jiang, and Kuncheng Li. Changes in hippocampal connectivity in the early stages of alzheimer's disease: Evidence from resting state fmri. *NeuroImage*, 31(2):496–504, 2006.

Larry Wasserman and John Lafferty. Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, 20:801–808, 2007.

Gary L. Wenk. Neuropathologic changes in Alzheimer's disease. *Journal of Clinical Psychiatry*, 64:7–10, 2003.

Jason Weston, Christina Leslie, Eugene Ie, Dengyong Zhou, Andre Elisseeff, and William Stafford Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005.

Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M. Thompson, and Jieping Ye. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102: 192–206, 2014.

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

Fei Xue, Rong Ma, and Hongzhe Li. Semi-supervised statistical inference for high-dimensional linear regression with blockwise missing data. *arXiv preprint arXiv:2106.03344*, 2021.

Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu. Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, 115(531):1406–1419, 2020.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Zhi-Hua Zhou and Ming Li. Semi-supervised regression with co-training. In *IJCAI*, pages 908–913, 2005.

Xiaojin Zhu and Andrew B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.