Statistica Sinica **34** (2024), 527-546
doi:https://doi.org/10.5705/ss.202021.0170

MULTI-RESPONSE REGRESSION FOR BLOCK-MISSING MULTI-MODAL DATA WITHOUT IMPUTATION

Haodong Wang, Quefeng Li and Yufeng Liu*

The University of North Carolina at Chapel Hill

Abstract: Multi-modal data are prevalent in many scientific fields. In this study, we consider the parameter estimation and variable selection for a multi-response regression using block-missing multi-modal data. Our method allows the dimensions of both the responses and the predictors to be large, and the responses to be incomplete and correlated, a common practical problem in high-dimensional settings. Our proposed method uses two steps to make a prediction from a multi-response linear regression model with block-missing multi-modal predictors. In the first step, without imputing missing data, we use all available data to estimate the covariance matrix of the predictors and the cross-covariance matrix between the predictors and the responses. In the second step, we use these matrices and a penalized method to simultaneously estimate the precision matrix of the response vector, given the predictors, and the sparse regression parameter matrix. Lastly, we demonstrate the effectiveness of the proposed method using theoretical studies, simulated examples, and an analysis of a multi-modal imaging data set from the Alzheimer's Disease Neuroimaging Initiative.

Key words and phrases: Inverse covariance matrix estimation, Lasso, missing data, moment estimation.

1. Introduction

With the prevalence of large-scale multi-modal data in various scientific fields, multi-response linear regression is attracting increasing attention in the statistics and machine learning communities (Rothman, Levina and Zhu (2010); Lee and Liu (2012); Loh and Zheng (2013)). Although linear regressions with a scalar response are well studied, many applications may have a vector as the response, for example, in biological problems (Kim and Xing (2012)). For example, for multi-tissue joint expression quantitative trait loci (eQTL) mapping (Molstad, Sun and Hsu (2020)), researchers predict gene expression values in multiple tissues simultaneously by using a weighted sum of eQTL genotypes. A separate prediction for each tissue is inefficient if the same genes in different tissues are correlated because of shared genetic variants or other unmeasured common regulators. In order to use data from all tissues simultaneously,

*Corresponding author.

Molstad, Sun and Hsu (2020) propose a joint eQTL model that considers cross-tissue expression dependence.

To apply variable selection methods to multi-response problems, one option is to separately fit each response using a single-response model. For example, the lasso is a well-studied variable selection method for single-response linear regression models (Tibshirani (1996)). However, although this is a straightforward method, it neglects the dependency structure between responses. Incorporating the dependency structure of the response vector enables us to obtain a more efficient multi-response linear regression approach in terms of estimation and prediction.

For multi-response regression problems, Breiman and Friedman (1997) proposed the curds and whey method to improve the prediction performance by using the dependencies between responses. Specifically, they first fit a single-response regression model for each response, and then modify the predicted values from these regressions by shrinking them using the canonical correlations between the response variables and the predictors. Another popular approach is to use dimension reduction. In particular, the reduced-rank regression (Izenman (1975)) minimizes the least squares criterion, subject to a constraint on the rank of the regression parameter matrix. Yuan et al. (2007) extended this method to include the high-dimensional settings, reducing the dimension by encouraging sparsity among the singular values of the parameter matrix. Nevertheless, although these methods achieve better prediction performance than when using a separate univariate regression, they do not address the problem of variable selection.

In order to handle correlated responses together with variable selection, we can estimate the precision matrix of the response vector, given the predictors, and the regression parameter matrix either separately or simultaneously (Lee and Liu (2012)). For a separate estimation, Cai et al. (2013) use a constrained ℓ_1 minimization that can be treated as a multivariate extension of the Dantzig selector to estimate the regression parameter matrix. After removing the regression effect using the estimated regression parameter matrix, the precision matrix of the error terms can be estimated accordingly. A potential drawback of this indirect method is that it ignores the relationships between the responses, given the predictors, when estimating the regression parameter matrix. Thus, in order to use all information more efficiently, it may be better to estimate the precision matrix and regression parameter matrix simultaneously. Existing joint estimation techniques include those of Rothman, Levina and Zhu (2010), Yin and Li (2011), and Lee and Liu (2012) who formulate the multi-response regression problem in a penalized log-likelihood framework to estimate the parameter and precision matrices simultaneously. Using a similar idea, Chen et al. (2018) propose an estimation procedure that estimates the parameter and precision matrices simultaneously based on the generalized Dantzig selector.

However, most existing multi-response linear regression methods deal only with complete data without missing entries, even though multi-modal data are often incomplete in practice. For instance, studies on Alzheimer's disease (AD) use data from different sources, including magnetic resonance imaging (MRI) of the brain, positron emission tomography (PET), and cerebrospinal fluid (CSF). In practice, observations of a certain modality can be missing completely, because patients drop out or other practical issues arise, leading to a block-wise missing data structure. Thus, it is important to integrate data from all modalities to improve model prediction and variable selection.

One way of handling incomplete multi-modal data is to simply remove observations with missing entries. However, this procedure may greatly reduce the number of observations and lead to loss of information. Another approach is to perform data imputation. However, existing imputation methods, such as matrix completion (Johnson (1990)) algorithms, may be unstable when the missing values occur in blocks. For such cases, Yu et al. (2020) proposed a direct sparse regression procedure using the covariance from the block-missing multi-modal data (DISCOM). They first use all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable, and then use these estimates and an extended Lasso-type estimator to estimate the coefficients. However, the DISCOM method considers only single-response regressions. Recently, Xue and Qu (2021) proposed the multiple block-wise imputation (MBI) method for a single-response regression when the data are block-wise missing. They developed an estimating equation approach to accommodate block-wise missing patterns in multi-modal data. The method is shown to have high selection accuracy and a low estimation error for a single-response regression with block-wise missing data. However, because their imputation method requires analyzing all combinations of blocks, it can be computationally expensive when the number of modalities is large.

Here, we consider a multi-response regression model for block-wise missing data. The main contribution of our method is to allow missing values in both the responses and the predictors, as well as correlations between responses. In contrast to most traditional methods, the proposed method can also be applied when no subject has complete observations. Our method includes two steps. The first step estimates each element of the covariance and cross-covariance matrices using all available observations without imputation. The second step uses a penalized approach to simultaneously estimate the sparse regression coefficient matrix and the precision matrix of the error terms. We show that this method exhibits estimation and model selection consistency in a high-dimensional setting. The results of our numerical studies and an analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data confirm that the proposed method performs competitively for block-wise missing data.

The remainder of the paper is organized as follows. In Section 2, we introduce the problem background and our model. In Section 3, we establish some theoretical properties of our proposed method, and in Sections 4 and 5, we present our simulation studies and a multi-modal ADNI data example, respectively.

2. Methodology

2.1. Problem setup and notation

Consider the following multi-response linear regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \mathcal{E}, \tag{2.1}$$

where $\mathbf{B}^* = (b_{jk}) \in \mathbb{R}^{p\times q}$ is an unknown $p \times q$ parameter matrix, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ is the $n \times q$ response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is the $n \times p$ design matrix, and $\mathcal{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$ is an $n \times q$ error matrix. We assume that $\{\mathbf{x}_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) realizations of a random vector $(X_1, \ldots, X_p)^\top$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_{XX} = (\sigma_{ij}^{XX}) \in \mathbb{R}^{p\times p}$. We use $\boldsymbol{\Sigma}_{XY} = (\sigma_{ij}^{XY}) \in \mathbb{R}^{p\times q}$ to denote the cross-covariance matrix between $\mathbf{x}_i$ and $\mathbf{y}_i$. We assume that the predictors come from multiple modalities, and there are p_k predictors in the kth modality. In addition, $\mathbf{X}$ has block-missing values. That is, for one sample, its measurements in one modality can be entirely missing. We assume the elements of $\mathbf{Y}$ can also be missing. The errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq})^\top$, for $i = 1, \ldots, n$, are i.i.d. realizations from a random vector $\boldsymbol{\epsilon}$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_\epsilon = (\sigma_{ij}^{EE}) \in \mathbb{R}^{q\times q}$. We let $\boldsymbol{C}^* = \boldsymbol{\Sigma}_\epsilon^{-1}$. Moreover, we assume $\mathbf{x}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated. Denote the support of $\mathbf{B}^*$ and $\mathbf{C}^*$ as $S_B = \{j : \mathrm{vec}(\mathbf{B}^*)_j \neq 0\}$ and $S_C = \{j : \mathrm{vec}(\mathbf{C}^*)_j \neq 0\}$, repectively, where "vec" denotes vectorization by a column operator. For a set S, we denote $|S|$ as its cardinality. Denote $s_B = |S_B|$, $s_C = |S_C|$, and $s = \max(s_B, s_C)$.

We employ the following notation throughout. The symbol $\mathbb{S}_+^{d\times d}$ denotes sets of $d \times d$ symmetric positive-definite matrices. For a square matrix $\mathbf{C} = (c_{ii'}) \in \mathbb{R}^{p\times p}$, we denote its trace as $\mathrm{tr}(\mathbf{C}) = \sum_i c_{ii}$ and its diagonal matrix as $\mathrm{diag}(\mathbf{C})$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we define its entrywise ℓ_1-norm as $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$, and its entrywise ℓ_∞-norm as $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$. In addition, we define its matrix ℓ_1-norm as $\|\mathbf{A}\|_{L_1} = \max_j \sum_i |a_{ij}|$, the matrix ℓ_∞-norm as $\|\mathbf{A}\|_{L_\infty} = \max_i \sum_j |a_{ij}|$, the spectral norm as $\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$, the Frobenius norm as $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, and the number of nonzero elements as $\|\mathbf{A}\|_0 = \sum_{i,j} \mathbb{I}(a_{ij} \neq 0)$. Denote the largest and smallest eigenvalues of $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. Denote the sub-matrix of $\mathbf{A}$ with row and column indices in I_1 and I_2 as $\mathbf{A}_{I_1 I_2}$. For a vector $\mathbf{v} \in \mathbb{R}^p$, denote $\mathbf{v}_{I_1}$ as the sub-vector of $\mathbf{v}$ with indices in I_1, $\|\mathbf{v}\|_1 = \sum_i |v_i|$, $\|\mathbf{v}\|_\infty = \max_i |v_i|$, $\|\mathbf{v}\|_{\min} = \min_i |v_i|$, and $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$. For a function $h(X)$, we use $\nabla_X h$ to denote a gradient or subgradient of h with respect to X, if it exists. Finally, we

write $a_n \lesssim b_n$ if $a_n \leq cb_n$ for some constant c, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2.2. Proposed multi-DISCOM method

If we separately apply a least squares estimation with the ℓ_1-norm penalty to each response, the multi-response linear regression model (2.1) essentially solves

$$\underset{\mathbf{B}}{\operatorname{argmin}}\, \mathbb{E}\left[\|\mathbf{Y}-\mathbf{XB}\|_F^2\right] + \lambda\|\mathbf{B}\|_1 = \underset{\mathbf{B}}{\operatorname{argmin}}\, \operatorname{tr}\left(\frac{1}{2}\mathbf{B}^\top \boldsymbol{\Sigma}_{XX}\mathbf{B} - \boldsymbol{\Sigma}_{XY}^\top \mathbf{B}\right) + \lambda\|\mathbf{B}\|_1, \tag{2.2}$$

where λ is a tuning parameter. We refer to this method as the separate lasso, with the solution denoted as $\hat{\mathbf{B}}^{LASSO}$. However, the approach fails to consider correlations between the responses, and may lead to poor predictive performance (see, e.g., Breiman and Friedman (1997)). To produce a better estimator, we propose incorporating $\boldsymbol{\Sigma}_\epsilon$ into the estimation of $\mathbf{B}^*$ and solving the following problem:

$$\hat{\mathbf{B}}^0 = \underset{\mathbf{B}}{\operatorname{argmin}}\, \operatorname{tr}\left[\mathbf{C}^*\hat{\boldsymbol{\Sigma}}_{\mathbf{YY}} + \mathbf{C}^*\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\mathbf{B} - \mathbf{2C}^*\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\right] + \lambda\|\mathbf{B}\|_1, \tag{2.3}$$

where λ is a tuning parameter, and $\hat{\boldsymbol{\Sigma}}_{YY}$, $\hat{\boldsymbol{\Sigma}}_{XX}$, and $\hat{\boldsymbol{\Sigma}}_{XY}$ are estimators of $\boldsymbol{\Sigma}_{YY}$, $\boldsymbol{\Sigma}_{XX}$, and $\boldsymbol{\Sigma}_{XY}$, respectively.

In practice, $\mathbf{C}^*$ is usually unknown. In case, we first estimate $\mathbf{C}^*$ using $\hat{\mathbf{C}}$, and then plug this into (2.3) and solve the following problem:

$$\hat{\mathbf{B}}^0 = \underset{\mathbf{B}}{\operatorname{argmin}}\, \operatorname{tr}\left[\hat{\mathbf{C}}\hat{\boldsymbol{\Sigma}}_{\mathbf{YY}} + \hat{\mathbf{C}}\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\mathbf{B} - \mathbf{2}\hat{\mathbf{C}}\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\right] + \lambda\|\mathbf{B}\|_1. \tag{2.4}$$

We refer to this method as the two-step weighted lasso. As shown in But as shown by the toy example in Section 2.2.1, the separate lasso may outperform this method in some problems.

We propose estimating $\mathbf{B}^*$ and $\mathbf{C}^*$ simultaneously by solving the following optimization problem:

$$\begin{aligned}(\hat{\mathbf{B}}, \hat{\mathbf{C}}) = \underset{\mathbf{C}\in\mathbb{S}_+^{q\times q},\mathbf{B}}{\operatorname{argmin}}\, & \operatorname{tr}\left[\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\mathbf{YY}} + \mathbf{C}\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\mathbf{B} - \mathbf{2}\mathbf{C}\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\right] \\ & + \lambda_B\|\mathbf{B}\|_1 + \lambda_C\|\mathbf{C}\|_1 - \log\det\mathbf{C},\end{aligned} \tag{2.5}$$

where λ_B and λ_C are tuning parameters. When λ_C is sufficiently large, Theorem 4 of Banerjee, El Ghaoui and d'Aspremont (2008) implies that all off-diagonal entries in $\hat{\mathbf{C}}$ become zero. Then, our proposed method (2.5) reduces to the separate lasso (2.2). For a univariate response regression problem, our proposed method (2.5) reduces to the DISCOM algorithm (Yu et al. (2020)). When there are no missing entries, (2.5) reduces to the sparse conditional Gaussian graphical model of Yin and Li (2011).

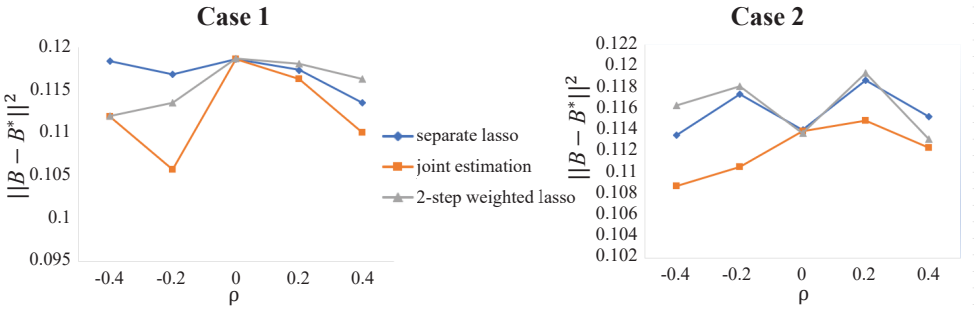

Figure 1. Plots of the estimation errors for the separated lasso, two-step weighted lasso and joint estimation when $\mathbf{\Sigma}_\epsilon = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$. The left panel is for $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 3.5 \end{smallmatrix}\right)$, and the right panel is for $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ -2 & 3.5 \end{smallmatrix}\right)$.

In the toy example in Section 2.2.1, our joint estimation model (2.5) outperforms the two-step weighted lasso and the separate lasso.

2.2.1. Toy example

For illustration, we consider a toy example similar to that in Lee and Liu (2012). Assume $p = q = 2$, $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, and $\mathbf{\Sigma}_\epsilon = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, where ρ is an unknown constant. We perform simulation studies for this example with 200 training samples, 300 tuning samples, and 1,000 testing samples. Set $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 3.5 \end{smallmatrix}\right)$ in Case 1, and $\left(\begin{smallmatrix} 0 & 0 \\ -2 & 3.5 \end{smallmatrix}\right)$ in Case 2. Figure 1 shows the estimation error for the separate lasso, two-step weighted lasso, and joint estimation model (2.5). In Case 1, the two-step weighted lasso has a smaller estimation error than that of the separate lasso when ρ is positive. The reverse is true when ρ is negative. In Case 2, the separate lasso has a smaller estimation error than that of the two-step weighted lasso when ρ is positive. The joint estimation model performs best in all cases.

The simulation results can be explained by the following calculations. With the penalty parameter λ, the solution of the separate lasso is given by $\hat{B}_{ij}^{\text{LASSO}} = \text{sign}(\hat{B}_{ij}^S)[\hat{B}_{ij}^S - \lambda/2]_+$, where $[u]_+ = u$ if $u \geq 0$, $[u]_+ = 0$ if $u < 0$, and $\hat{\mathbf{B}}^S = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.

We can show that the two-step weighted lasso (2.4) is equivalent to

$$\hat{\mathbf{B}}^{2step} = \underset{\mathbf{B}}{\text{argmin}} \left[(\text{vec}(\mathbf{B}) - \text{vec}(\mathbf{B}^S))^\top (\mathbf{I}_2 \otimes \hat{\mathbf{C}})(\text{vec}(\mathbf{B}) - \text{vec}(\mathbf{B}^S)) + \| \text{vec}(\mathbf{B})\|_1 \right]. \tag{2.6}$$

When the estimate $\hat{\mathbf{C}}$ is accurate, $\hat{\mathbf{B}}^{2step}$ should be very close to the solution of (2.3), where we use $\mathbf{\Sigma}_\epsilon^{-1}$ as the weight. After we plug $\hat{\mathbf{C}} = \mathbf{\Sigma}_\epsilon^{-1}$ into (2.6), the solution is given by $\hat{B}_{ij}^{2step} = \text{sign}(\hat{B}_{ij}^S)[|\hat{B}_{ij}^S| - \lambda(1+\rho)/2]_+$ when $\text{sign}(\hat{B}_{i1}^S\hat{B}_{i2}^S) = 1$, and $\hat{B}_{ij}^{2step} = \text{sign}(\hat{B}_{ij}^S)[|\hat{B}_{ij}^S| - \lambda(1-\rho)/2]_+$ when $\text{sign}(\hat{B}_{i1}^S\hat{B}_{i2}^S) = -1$. Compared with $\hat{B}_{ij}^{\text{LASSO}} = \text{sign}(\hat{B}_{ij}^S)[\hat{B}_{ij}^S - \lambda/2]_+$, $\hat{B}_{ij}^{2\text{step}}$ differs only in the shrinkage amount for each entry. The shrinkage amounts for all entries of the separate lasso are

the same, and depend only on the tuning parameter λ. The shrinkage amounts for all entries of the two-step weighted lasso depend on ρ, λ, and the sign of $\hat{\mathbf{B}}^S$. Each entry of the two-step weighted lasso may have different shrinkage amounts.

We consider two cases of ρ in Case 1, where $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 3.5 \end{smallmatrix}\right)$. Because B^*_{21} and B^*_{22} are far from zero, for simplicity, we assume that $\text{sign}(\hat{B}^S_{21}) = \text{sign}(\hat{B}^S_{22}) = 1$.

1. Consider $\rho = -0.4$. When $\text{sign}(\hat{B}^S_{11}\hat{B}^S_{12}) = -1$, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are 0.7λ, and those for $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ are 0.3λ. Thus, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are smaller than those for $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$. Therefore, with the tuning parameter λ that shrinks $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ to zero, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are also smaller than those for $\hat{B}^{LASSO}_{21}$ and $\hat{B}^{LASSO}_{22}$. Thus, the two-step weighted lasso has a smaller estimation error than that of the separate lasso in this scenario. When $\text{sign}(\hat{B}^S_{11}\hat{B}^S_{12}) = 1$, the shrinkage amounts for all entries in $\hat{\mathbf{B}}^{2step}$ are equal.

2. Consider $\rho = 0.4$. When $\text{sign}(\hat{B}^S_{11}\hat{B}^S_{12}) = -1$, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are 0.3λ, and those for $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ are 0.7λ. Therefore, with the tuning parameter λ that shrinks $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ to zero, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are larger than those for $\hat{B}^{LASSO}_{21}$ and $\hat{B}^{LASSO}_{22}$. Thus, the separate lasso is preferred to the two-step weighted lasso in this scenario. When $\text{sign}(\hat{B}^S_{11}\hat{B}^S_{12}) = 1$, all entries in $\hat{\mathbf{B}}^{2step}$ have the same shrinkage amount.

In Case 2, where $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ -2 & 3.5 \end{smallmatrix}\right)$, the two-step weighted lasso is preferred to the separate lasso only when ρ is negative. In conclusion, the performance of the two-step weighted lasso compared with that of the separate lasso depends on the sign of $\mathbf{B}^*$ and the covariance matrix $\mathbf{\Sigma}_\epsilon$. In contrast, the joint estimation model (2.5) is more flexible. When $\mathbf{\Sigma}_\epsilon$ and $\mathbf{B}^*$ favor the separate lasso, the joint estimation model (2.5) performs better by choosing a large λ_C. Otherwise, it can perform better by choosing a relatively small λ_C, and thus performs competitively in all cases.

2.2.2. Covariance estimation

Now, we show how to obtain $\hat{\mathbf{\Sigma}}_{XX}$, $\hat{\mathbf{\Sigma}}_{XY}$, and $\hat{\mathbf{\Sigma}}_{YY}$ when the data exhibit block-missing values. The following notation is used throughout. For the jth predictor, define $S^X_j = \{i : x_{ij} \text{ is not missing}\}$. For the jth response, define $S^Y_j = \{i : y_{ij} \text{ is not missing}\}$. Define $S^{XX}_{jk} = \{i : x_{ij} \text{ and } x_{ik} \text{ are not missing}\}$, $S^{XY}_{jk} = \{i : x_{ij} \text{ and } y_{ik} \text{ are not missing}\}$, $S^{XX/Y}_{jkl} = \{i : x_{ij}, x_{ik} \text{ are not missing, but } y_{il} \text{ is missing}\}$, $S^{XY/X}_{jkl} = \{i : x_{ij}, y_{ik} \text{ are not missing but } x_{il} \text{ is missing}\}$, and $S^{YY}_{jk} = \{i : y_{ij} \text{ and } y_{ik} \text{ are not missing}\}$. Denote the cardinality of S^X_j, S^Y_j, S^{XX}_{jk}, S^{XY}_{jk}, $S^{XX/Y}_{jkl}$, $S^{XY/X}_{jkl}$, and S^{YY}_{jk} as n^X_j, n^Y_j n^{XX}_{jk}, n^{XY}_{jk}, $n^{XX/Y}_{jkl}$, $n^{XY/X}_{jkl}$, and n^{YY}_{jk}, respectively. Denote $n_X = \min_j |S^X_j|$, $n_{XX} = \min_{j,k} |S^{XX}_{jk}|$, $n_{XY} = \min_{j,k} |S^{XY}_{jk}|$,

$n_{YY} = \min_{j,k} |S_{jk}^{YY}|$, $n_{XX/Y} = \max_{j,k,l} |S_{jkl}^{XX/Y}|$, and $n_{XY/X} = \max_{j,k,l} |S_{jkl}^{XY/X}|$.

We propose using the initial estimators of $\mathbf{\Sigma}_{XX}$, $\mathbf{\Sigma}_{XY}$, and $\mathbf{\Sigma}_{YY}$ as the sample covariance matrices from all available data, that is, $\tilde{\mathbf{\Sigma}}_{XX} = (\tilde{\sigma}_{jt}^{XX}), \tilde{\mathbf{\Sigma}}_{XY} = (\tilde{\sigma}_{jt}^{XY}), \hat{\mathbf{\Sigma}}_{YY} = (\hat{\sigma}_{jt}^{YY})$, where $\tilde{\sigma}_{jt}^{XX} = \sum_{i \in S_{jt}^{XX}} x_{ij} x_{it} / n_{jt}^{XX}$, $\tilde{\sigma}_{jt}^{XY} = \sum_{i \in S_{jt}^{XY}} x_{ij} y_{it} / n_{jt}^{XY}$, and

$$\hat{\sigma}_{jt}^{YY} = \frac{1}{n_{jt}^{YY}} \sum_{i \in S_{jt}^{YY}} y_{ij} y_{it}. \tag{2.7}$$

Note that our method requires that $\tilde{\mathbf{\Sigma}}_{XX}$, $\tilde{\mathbf{\Sigma}}_{XY}$, and $\hat{\mathbf{\Sigma}}_{YY}$ be unbiased estimators of their counterparts. When the missingness in $\mathbf{X}$ and Y is completely at random, the unbiasedness assumption is satisfied. However, this assumption may also hold under other missing mechanisms. For our theory, we do not specify any particular missing mechanism, and the unbiasedness assumption suffices.

For block-missing data $\mathbf{X}$, the estimate $\tilde{\mathbf{\Sigma}}_{XX}$ can be ill-conditioned and have negative eigenvalues. Therefore, it may not be a good estimate of $\mathbf{\Sigma}_{XX}$, and cannot be used in (2.5) directly. Next, we introduce an estimator that is both well conditioned and more accurate than the initial estimate $\tilde{\mathbf{\Sigma}}_{XX}$. According to the partition of the predictors into K modalities, $\tilde{\mathbf{\Sigma}}_{XX}$ can be partitioned into K^2 blocks, denoted by $\tilde{\mathbf{\Sigma}}^{k_1 k_2}$, for $1 \le k_1, k_2 \le K$, where $\tilde{\mathbf{\Sigma}}^{k_1 k_2}$ is a $p_{k_1} \times p_{k_2}$ matrix. We denote

$$\tilde{\mathbf{\Sigma}}_I = \begin{pmatrix} \tilde{\mathbf{\Sigma}}^{11} & & & \\ & \tilde{\mathbf{\Sigma}}^{22} & & \\ & & \ddots & \\ & & & \tilde{\mathbf{\Sigma}}^{KK} \end{pmatrix} \quad \text{and} \quad \tilde{\mathbf{\Sigma}}_C = \begin{pmatrix} \mathbf{0} & \tilde{\mathbf{\Sigma}}^{12} & \dots & \tilde{\mathbf{\Sigma}}^{1K} \\ \tilde{\mathbf{\Sigma}}^{21} & \mathbf{0} & \dots & \tilde{\mathbf{\Sigma}}^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{\Sigma}}^{K1} & \tilde{\mathbf{\Sigma}}^{K2} & \dots & \mathbf{0} \end{pmatrix},$$

where $\tilde{\mathbf{\Sigma}}_I$ is called the intra-modality sample covariance matrix and is a $p \times p$ block-diagonal matrix containing K diagonal blocks of $\tilde{\mathbf{\Sigma}}_{XX}$, and $\tilde{\mathbf{\Sigma}}_C = \tilde{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}_I$ is called the cross-modality sample covariance matrix containing all off-diagonal blocks of $\tilde{\mathbf{\Sigma}}_{XX}$. Let $\mathbf{\Sigma}_I$ and $\mathbf{\Sigma}_C$ be the true intra-modality and cross-modality covariance matrices, respectively. For block-missing multi-modal data, the imbalanced sample sizes mean that the estimate $\tilde{\mathbf{\Sigma}}_I$ can be relatively accurate, while the estimate $\tilde{\mathbf{\Sigma}}_C$ can be inaccurate. In that case, we estimate $\mathbf{\Sigma}_{XX}$ using a linear combination of $\tilde{\mathbf{\Sigma}}_I$ and $\tilde{\mathbf{\Sigma}}_C$ with different weights. In addition, to ensure the positive definiteness of our estimation, we adopt the idea of a shrinkage estimation of the covariance matrix (Fisher and Sun (2011)) and add the diagonal matrix $\operatorname{diag}(\tilde{\mathbf{\Sigma}}_I)$ to our estimator,

$$\hat{\mathbf{\Sigma}}_{XX} = \alpha_1 \tilde{\mathbf{\Sigma}}_I + (1 - \alpha_1) \operatorname{diag}(\tilde{\mathbf{\Sigma}}_I) + \alpha_2 \tilde{\mathbf{\Sigma}}_C, \tag{2.8}$$

where $\alpha_1, \alpha_2 \in [0, 1]$ are two shrinkage weights. We add the diagonal matrix $\operatorname{diag}(\tilde{\mathbf{\Sigma}}_I)$ to ensure the diagonal entries of our estimator are not shrunk.

By Weyl's theorem, the eigenvalues of our estimator are greater than or equal to $\alpha_1\lambda_{\min}(\tilde{\mathbf{\Sigma}}_I)+(1-\alpha_1)\lambda_{\min}(\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I))+\alpha_2\lambda_{\min}(\tilde{\mathbf{\Sigma}}_C)$. Because $\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)$ is a positive-definite matrix, we can guarantee that the eigenvalues of our estimator are positive by carefully selecting the tuning parameters α_1 and α_2.

As dicussed previously, our estimator $\hat{\mathbf{\Sigma}}_{XX}$ is a shrinkage estimator. Using a similar idea, we use a shrinkage estimator to estimate $\mathbf{\Sigma}_{XY}$. That is, we propose estimating $\mathbf{\Sigma}_{XY}$ by

$$\hat{\mathbf{\Sigma}}_{XY}=\alpha_3\tilde{\mathbf{\Sigma}}_{XY}, \tag{2.9}$$

where $\alpha_3\in[0,1]$ is the shrinkage weight. Here, we want to find the optimal linear combination $\hat{\mathbf{\Sigma}}^*_{XY}=\alpha_3^*\tilde{\mathbf{\Sigma}}_{XY}$ that minimizes the expected quadratic loss $\mathbb{E}\|\hat{\mathbf{\Sigma}}^*_{XY}-\mathbf{\Sigma}_{XY}\|_F$.

Here, we consider only a relative low dimension of Y, with not too many incomplete observations, so we use $\hat{\mathbf{\Sigma}}_{YY}$ defined in (2.7) directly. However, when the dimension of Y is very high or there are many incomplete observations of Y, a shrinkage estimator of $\mathbf{\Sigma}_{YY}$ is recommended instead.

Denote $\gamma^*=(\gamma_1^*,\ldots,\gamma_K^*)^\top=(\mathrm{tr}(\mathbf{\Sigma}^{11})/p_1,\ldots,\mathrm{tr}(\mathbf{\Sigma}^{KK})/p_K)^\top$, $\delta_I=\sqrt{\mathbb{E}\|\tilde{\mathbf{\Sigma}}_I-\mathbf{\Sigma}_I\|_F^2}$, $\delta_C=\sqrt{\mathbb{E}\|\tilde{\mathbf{\Sigma}}_C-\mathbf{\Sigma}_C\|_F^2}$, $\delta_{XY}=\sqrt{\mathbb{E}\|\tilde{\mathbf{\Sigma}}_{XY}-\mathbf{\Sigma}_{XY}\|_F^2}$ and $\theta=\|\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)-\mathbf{\Sigma}_I\|_F$. The optimal choice for the weights of α_1,α_2, and α_3 is stated in Proposition 1.

Proposition 1. *The solutions to the two optimization problems*

$$(\alpha_1^*,\alpha_2^*)=\underset{\alpha_1,\alpha_2}{\mathrm{argmin}}\,\mathbb{E}\|\hat{\mathbf{\Sigma}}_{XX}-\mathbf{\Sigma}_{XX}\|_F^2 \tag{2.10}$$

$$\alpha_3^*=\underset{\alpha_3}{\mathrm{argmin}}\,\mathbb{E}\|\hat{\mathbf{\Sigma}}_{XY}-\mathbf{\Sigma}_{XY}\|_F^2 \tag{2.11}$$

are

$$\alpha_1^*=\frac{\theta^2}{\theta^2+\delta_I^2},\quad \alpha_2^*=\frac{\|\mathbf{\Sigma}_C\|_F^2}{\|\mathbf{\Sigma}_C\|_F^2+{\delta_C}^2},\ \textit{and}\ \alpha_3^*=\frac{\|\mathbf{\Sigma}_{XY}\|_F^2}{\|\mathbf{\Sigma}_{XY}\|_F^2+{\delta_{XY}}^2}.$$

In addition, for $\hat{\mathbf{\Sigma}}^*_{XX}=\alpha_1^*\tilde{\mathbf{\Sigma}}_I+(1-\alpha_1^*)\,\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)+\alpha_2^*\tilde{\mathbf{\Sigma}}_C$ *and* $\hat{\mathbf{\Sigma}}^*_{XY}=\alpha_3^*\tilde{\mathbf{\Sigma}}_{XY}$, *we have*

$$\mathbb{E}\left\|\hat{\mathbf{\Sigma}}^*_{XX}-\mathbf{\Sigma}_{XX}\right\|_F^2=\frac{\delta_I^2\theta^2}{\delta_I^2+\theta^2}+\frac{{\delta_C}^2\|\mathbf{\Sigma}_C\|_F^2}{{\delta_C}^2+\|\mathbf{\Sigma}_C\|_F^2}\le\delta_I^2+{\delta_C}^2=\mathbb{E}\|\tilde{\mathbf{\Sigma}}_{XX}-\mathbf{\Sigma}_{XX}\|_F^2,$$

$$\mathbb{E}\left\|\hat{\mathbf{\Sigma}}^*_{XY}-\mathbf{\Sigma}_{XY}\right\|_F^2=\frac{{\delta_{XY}}^2\|\mathbf{\Sigma}_{XY}\|_F^2}{{\delta_{XY}}^2+\|\mathbf{\Sigma}_{XY}\|_F^2}\le{\delta_{XY}}^2=\mathbb{E}\|\tilde{\mathbf{\Sigma}}_{XY}-\mathbf{\Sigma}_{XY}\|_F^2.$$

Define the ℓ_2-error of the estimators $\hat{\mathbf{\Sigma}}_{XX}$ and $\hat{\mathbf{\Sigma}}_{XY}$ as $\mathbb{E}\|\hat{\mathbf{\Sigma}}_{XX}-\mathbf{\Sigma}_{XX}\|_F^2$ and $\mathbb{E}\|\hat{\mathbf{\Sigma}}_{XY}-\mathbf{\Sigma}_{XY}\|_F^2$, respectively. Proposition 1 shows that our estimator is more accurate than the sample covariance matrix.

Proposition 1 is closely related to Proposition 1 of Yu et al. (2020). They calculated the optimal weight and estimation error for their proposed estimator

$\hat{\mathbf{\Sigma}}^{*}_{XX,DISCOM}$ of $\mathbf{\Sigma}_{XX}$, where the estimation error is

$$\mathbb{E}\|\hat{\mathbf{\Sigma}}_{XX,DISCOM} - \mathbf{\Sigma}_{XX}\|_F^2 = \frac{\delta_I^2\tilde{\theta}^2}{\delta_I^2 + \tilde{\theta}^2} + \frac{{\delta_C}^2\,\|\mathbf{\Sigma}_C\|_F^2}{{\delta_C}^2 + \|\mathbf{\Sigma}_C\|_F^2},$$

and $\tilde{\theta}^2 = \|\mathrm{tr}(\mathbf{\Sigma})\mathbf{I_p}/p - \mathbf{\Sigma}_I\|_F^2$. Here, our estimator $\hat{\mathbf{\Sigma}}_{XX}$ has a smaller ℓ_2-error than that of their estimator, and our weighted estimator $\hat{\mathbf{\Sigma}}_{XY}$ is more accurate than the sample covariance matrix.

2.3. Computational algorithm

In this section, we describe the computational algorithm used to solve the optimization problem (2.5). Because (2.5) is a bi-convex problem, the standard approach to solving it is to use the alternating minimization method. In particular, starting with some given initial point $(\hat{\mathbf{B}}_0, \hat{\mathbf{C}}_0)$ at the tth iteration, we solve solving the following problems:

$$\begin{aligned}\hat{\mathbf{B}}_t = &\ \underset{\mathbf{B}}{\mathrm{argmin}}\,\mathrm{tr}\left[\hat{\mathbf{C}}_{t-1}\hat{\mathbf{\Sigma}}_{YY} + \hat{\mathbf{C}}_{t-1}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\hat{\mathbf{C}}_{t-1}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{XY}\right] \\ &+ \lambda_B\|\mathbf{B}\|_1, \qquad (2.12)\\ \hat{\mathbf{C}}_t = &\ \underset{\mathbf{C}\in\mathbb{S}_+^{q\times q}}{\mathrm{argmin}}\,\mathrm{tr}\left[\mathbf{C}\hat{\mathbf{\Sigma}}_{YY} + \mathbf{C}\hat{\mathbf{B}}_{t-1}^\top\hat{\mathbf{\Sigma}}_{XX}\hat{\mathbf{B}}_{t-1} - 2\mathbf{C}\hat{\mathbf{B}}_{t-1}^\top\hat{\mathbf{\Sigma}}_{XY}\right] \\ &+ \lambda_C\|\mathbf{C}\|_1 - \log\det\mathbf{C}. \qquad (2.13)\end{aligned}$$

In each iteration of our algorithm, given $\hat{\mathbf{C}}_{t-1}$, we first update the estimator $\hat{\mathbf{B}}_t$ by solving (2.12). Because (2.12) is quadratic in $\mathbf{B}$, we use the coordinate descent algorithm to solve it. Then, we adopt the graphical lasso method of Friedman, Hastie and Tibshirani (2008) to solve (2.13). We summarize the above procedures in Algorithm 1.

3. Theoretical Study

We establish the following theoretical results. First, we prove in Theorem 1 that the proposed estimators $\hat{\mathbf{\Sigma}}_{XX}$, $\hat{\mathbf{\Sigma}}_{XY}$ and $\hat{\mathbf{\Sigma}}_{YY}$ are consistent with high probability. We then show the convergence rate of our proposed estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ in Theorem 2. Finally, the selection consistency of our proposed method is shown in Theorem 3. The technical assumptions (A1) to (A5), and all proofs are provided in the Supplementary Material. In the following analysis, we allow p and q to diverge as n_{XX}, n_{XY} and n_{YY} increase.

In Theorem 1, we prove the large deviation bounds for our proposed estimators $\hat{\mathbf{\Sigma}}_{XX}$, $\hat{\mathbf{\Sigma}}_{XY}$ and $\hat{\mathbf{\Sigma}}_{YY}$.

Theorem 1. *Suppose* $1-\alpha_1 = O(\sqrt{\log p/n_X})$, $1-\alpha_2 = O(\sqrt{\log p/n_{XX}})$, *and* $1-\alpha_3 = O(\sqrt{\log(pq)/n_{XY}})$. *If Conditions* (A1) *and* (A2) *hold, there exists*

Algorithm 1: Alternating minimization updating algorithm.

Input: $\mathbf{X}$, $\mathbf{Y}$, λ_C, λ_B

Output: $\hat{\mathbf{B}}, \hat{\mathbf{C}}$

Obtain $\hat{\mathbf{\Sigma}}_{XX}$ by (2.8), $\hat{\mathbf{\Sigma}}_{XY}$ by (2.9), $\hat{\mathbf{\Sigma}}_{YY}$ by (2.7).

Initialize with

$$\hat{\mathbf{B}}_0 = \underset{\mathbf{B}}{\operatorname{argmin}} \operatorname{tr}\left[\hat{\mathbf{\Sigma}}_{YY} + \mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_{B_0}\|\mathbf{B}\|_1, \tag{3.4}$$

$$\hat{\mathbf{C}}_0 = \underset{\|\mathbf{C}\|_1 \le R, \mathbf{C}\in\mathbb{S}_+^{d\times d}}{\operatorname{argmin}} \operatorname{tr}(\mathbf{C}\hat{\mathbf{\Sigma}}_0) - \log\det(\mathbf{C}) + \lambda_{C_0}\|\mathbf{C}\|_1, \tag{3.5}$$

where R is a large enough tuning parameter which is usually chosen to be $\lambda_{C_0}^{-1}$ (Loh and Wainwright (2015)) and $\hat{\mathbf{\Sigma}}_0 = \hat{\mathbf{\Sigma}}_{YY} - 2\hat{\mathbf{\Sigma}}_{XY}^\top \hat{\mathbf{B}}_0 + \hat{\mathbf{B}}_0^\top \hat{\mathbf{\Sigma}}_{XX}\hat{\mathbf{B}}_0$.

while $\max\{\|\hat{\mathbf{B}}_t - \hat{\mathbf{B}}_{t-1}\|_F, \|\hat{\mathbf{C}}_t - \hat{\mathbf{C}}_{t-1}\|_F\} >$ *threshold* **do**

For a given $\hat{\mathbf{C}}_{t-1}$, let

$$\hat{\mathbf{B}}_t = \underset{\mathbf{B}}{\operatorname{argmin}} \operatorname{tr}\left[\hat{\mathbf{C}}_{t-1}\hat{\mathbf{\Sigma}}_{YY} + \hat{\mathbf{C}}_{t-1}\mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\hat{\mathbf{C}}_{t-1}\mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_B\|\mathbf{B}\|_1;$$

For a given $\hat{\mathbf{B}}_t$, let

$$\hat{\mathbf{C}}_t = \underset{\|\mathbf{C}\|_1 \le R, \mathbf{C}\in\mathbb{S}_+^{q\times q}}{\operatorname{argmin}} \operatorname{tr}\left[\mathbf{C}\hat{\mathbf{\Sigma}}_{YY} + \mathbf{C}\hat{\mathbf{B}}_{t-1}^\top \hat{\mathbf{\Sigma}}_{XX}\hat{\mathbf{B}}_{t-1} - 2\mathbf{C}\hat{\mathbf{B}}_{t-1}^\top \hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_C\|\mathbf{C}\|_1 - \log\det\mathbf{C},$$

return $\hat{\mathbf{C}}_t$, $\hat{\mathbf{B}}_t$.

positive constants v_1', v_2', and v_3' such that

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\right\|_\infty \ge v_1'\sqrt{\frac{\log p}{n_{XX}}}\right) \le \frac{4}{p}, \tag{3.1}$$

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY}\right\|_\infty \ge v_2'\sqrt{\frac{\log(pq)}{n_{XY}}}\right) \le \frac{4}{pq}, \tag{3.2}$$

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{YY} - \mathbf{\Sigma}_{YY}\right\|_\infty \ge v_3'\sqrt{\frac{\log q}{n_{YY}}}\right) \le \frac{4}{q}. \tag{3.3}$$

If we only use samples with complete observations, sample covariance estimators $\tilde{\mathbf{\Sigma}}_{XX,\text{complete}}$, $\tilde{\mathbf{\Sigma}}_{XX,\text{complete}}$ and $\tilde{\mathbf{\Sigma}}_{XX,\text{complete}}$ have the following convergence rates

$$\left\|\tilde{\mathbf{\Sigma}}_{XX,\text{complete}} - \mathbf{\Sigma}_{XX}\right\|_\infty = O_p\left(\sqrt{\frac{\log p}{n_{\text{complete}}}}\right),$$

$$\left\|\tilde{\boldsymbol{\Sigma}}_{XY,\text{complete}} - \boldsymbol{\Sigma}_{XY}\right\|_\infty = O_p\left(\sqrt{\frac{\log(pq)}{n_{\text{complete}}}}\right),$$
$$\left\|\tilde{\boldsymbol{\Sigma}}_{YY,\text{complete}} - \boldsymbol{\Sigma}_{YY}\right\|_\infty = O_p\left(\sqrt{\frac{\log q}{n_{\text{complete}}}}\right),$$

where n_{complete} is the number of samples with complete observations; see Yu et al. (2020). For block-missing data, n_{complete} can be much smaller than n_{XX}, n_{XY} and n_{YY}.

Next, we give the properties of initial estimators $\hat{\mathbf{B}}_0$ and $\hat{\mathbf{C}}_0$. The following lemma describes estimation consistency of the initial estimator $\hat{\mathbf{B}}_0$.

Lemma 1. *Suppose Conditions* (A1)–(A4) *hold,* $1-\alpha_1 = O(\sqrt{\log p/n_X})$, $1-\alpha_2 = O(\sqrt{\log p/n_{XX}})$, *and* $1-\alpha_3 = O(\sqrt{\log(pq)/n_{XY}})$. *If we choose* $\lambda_{B_0} = C(\log(pq)/\min(n_{XY}, n_{XX}))^{1/2}\|\mathbf{B}^*\|_{L_1}$ *for some large enough constant* C, *then with probability at least* $1-4/p-4/(pq)$, *the initial estimator* $\hat{\mathbf{B}}_0 = \operatorname{argmin}_{\mathbf{B}} \operatorname{tr}[\hat{\boldsymbol{\Sigma}}_{YY} + \mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{XX}\mathbf{B} - 2\mathbf{B}^\top\hat{\boldsymbol{\Sigma}}_{XY}] + \lambda_B\|\mathbf{B}\|_1$ *satisfies*

$$\begin{aligned}\left\|\hat{\mathbf{B}}_0 - \mathbf{B}^*\right\|_F &\lesssim \sqrt{qs_B}\left\|\hat{\boldsymbol{\Sigma}}_{XY} - \hat{\boldsymbol{\Sigma}}_{XX}\mathbf{B}^*\right\|_\infty \\ &\lesssim \|\mathbf{B}^*\|_{L_1}\sqrt{\frac{qs_B\log(pq)}{\min(n_{XX}, n_{XY})}}.\end{aligned}$$

Cai et al. (2013) showed that when there is no missing data and the true coefficient $\mathbf{B}^*$ is exactly sparse, their estimator $\hat{\mathbf{B}}_{Cai}$ has the convergence rate of $\|\hat{\mathbf{B}}_{Cai} - \mathbf{B}^*\|_F = O_p(N_p\sqrt{qs_B\log(pq)/n})$, where n is the sample size of the data and N_p is the upper bound of $\|\boldsymbol{\Sigma}_{XX}^{-1}\|_{L_\infty}$. When there is no missing data, our initial estimator $\hat{\mathbf{B}}_0$ has the convergence rate of $\|\hat{\mathbf{B}}_0 - \mathbf{B}^*\|_F = O_p(\|\mathbf{B}^*\|_{L_1}\sqrt{qs_B\log(pq)/n})$. If we assume $\|\mathbf{B}^*\|_{L_1} \asymp \|\boldsymbol{\Sigma}_{XX}^{-1}\|_{L_\infty}$, the convergence rate of $\hat{\mathbf{B}}_0$ is the same as that of $\hat{\mathbf{B}}_{Cai}$. When the data are block-wise missing, and we only use complete samples to estimate $\mathbf{B}^*$, we will have $\|\hat{\mathbf{B}}_0 - \mathbf{B}^*\|_F = O_p(\|\mathbf{B}^*\|_{L_1}\sqrt{qs_B\log(pq)/n_{\text{complete}}})$, which can be much slower than the rate in Lemma 1 as n_{complete} is typically much smaller than n_{XX} and n_{XY} for block-wise missing data.

For the single-response regression with block-wise missing data, the result in Lemma 1 is the same as Theorem 2 in Yu et al. (2020) and the estimator $\hat{\mathbf{B}}_0$ performs well when the dimension of $\mathbf{Y}$ is small. But when the dimension of $\mathbf{Y}$ becomes large, the estimator $\hat{\mathbf{B}}_0$ may perform poorly.

The following lemma describes consistency of our initial estimator $\hat{\mathbf{C}}_0$.

Lemma 2. *Suppose Conditions* (A1)–(A4) *hold,* $1-\alpha_1 = O(\sqrt{\log p/n_X})$, $1-\alpha_2 = O(\sqrt{\log p/n_{XX}})$, $1-\alpha_3 = O(\sqrt{\log(pq)/n_{XY}})$. *If we choose* $\lambda_{C_0} = C\|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}\left(\|\mathbf{B}^*\|_{L_1} + s_B\sqrt{q}\right)(\log(pq)/\min(n_{XX}, n_{XY}))^{1/2}$ *for a large enough* C, *it holds with probability at least* $1-4/p-4/(pq)-4/q$ *that*

$$\left\|\hat{\mathbf{C}}_0 - \mathbf{C}^*\right\|_F \lesssim \sqrt{s_C}\|\mathbf{C}^*\|_2^2\|\mathbf{\Sigma}_\epsilon - \hat{\mathbf{C}}_0^{-1}\|_\infty$$
$$\lesssim \|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}(\|\mathbf{B}^*\|_{L_1} + s_B\sqrt{q})\sqrt{\frac{s_C\log(pq)}{\min(n_{XX}, n_{XY})}}.$$

There are two terms in the estimation error bound of $\hat{\mathbf{C}}_0$. The first term $\|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}^2\sqrt{(s_C\log(pq))/\min(n_{XX}, n_{XY})}$ comes from the error induced by using incomplete observations to estimate $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{XY}$. The second term $\|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}s_B\sqrt{(s_Cq\log(pq))/\min(n_{XX}, n_{XY})}$ comes from the estimation error of $\hat{\mathbf{B}}_0$.

We next derive the convergence rates of $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$. The convergence rates are related to $n_{XX/Y}$ and $n_{XY/X}$, which are fractions of n_{XX} and n_{XY} respectively. Hence, we let $n_{XX/Y} \asymp n_{XX}^{\tau_1}$ and $n_{XY/X} \asymp n_{XY}^{\tau_2}$ with $\tau_1, \tau_2 \in \{-\infty\} \cup [0, 1]$. When the responses are complete while the covariates have missing entries, $n_{XX/Y} = 0$ and $\tau_1 = -\infty$, $n_{XY/X} > 0$ and $\tau_2 \in [0, 1]$. When the covariates are complete while the responses have missing entries, $n_{XY/X} = 0$ and $\tau_2 = -\infty$, $n_{XX/Y} > 0$ and $\tau_1 \in [0, 1]$. When both the responses and covaraites are complete, $n_{XX/Y} = n_{XY/X} = 0$ and $\tau_1 = \tau_2 = -\infty$. Theorem 2 below establishes the consistency of proposed estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ in (2.5).

Theorem 2. *Suppose Conditions* (A1)–(A4) *hold,* $1 - \alpha_1 = O(\sqrt{\log p/n_X})$, $1 - \alpha_2 = O(\sqrt{\log p/n_{XX}})$, $1 - \alpha_3 = O(\sqrt{\log(pq)/n_{XY}})$. *If we choose* λ_B *and* λ_C *satisfying* $\lambda_B = C((\log p)^{1/2}/\min(n_{XX}^{1-\tau_1/2}, n_{XY}^{1-\tau_2/2})\|\mathbf{B}^*\mathbf{C}^*\|_{L_1} + \{\log(pq)/n_{XY}\}^{1/2})$ *and* $\lambda_C = C\|\mathbf{C}^*\|_2^2[\|\mathbf{B}^*\|_{L_1}^2 + s_B\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}/\min(n_{XX}^{1/2-\tau_1/2}, n_{XY}^{1/2-\tau_2/2})](\log(pq)/\min(n_{XX}, n_{XY}))^{1/2}$ *for a large enough* C, *then it holds with probability at least* $1 - 4/p - 4/(pq) - 4/q$ *that*

$$\left\|\hat{\mathbf{B}} - \mathbf{B}^*\right\|_F \lesssim \sqrt{s_B}\left(\frac{\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{\min\left(n_{XX}^{1-\tau_1/2}, n_{XY}^{1-\tau_2/2}\right)} + \left\{\frac{\log(pq)}{n_{XY}}\right\}^{1/2}\right),$$

$$\left\|\hat{\mathbf{C}} - \mathbf{C}^*\right\|_F \lesssim \sqrt{s_C}\|\mathbf{C}^*\|_2^2\left(\frac{s_B\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{\min\left(n_{XX}^{1-\tau_1/2}, n_{XY}^{1-\tau_2/2}\right)} + \frac{\|\mathbf{B}^*\|_{L_1}^2(\log(pq))^{1/2}}{\min\left(n_{XX}^{1/2}, n_{XY}^{1/2}\right)}\right)$$

$$\left\|\hat{\mathbf{B}} - \mathbf{B}^*\right\|_1 \lesssim s_B\left(\frac{\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{\min\left(n_{XX}^{1-\tau_1/2}, n_{XY}^{1-\tau_2/2}\right)} + \left\{\frac{\log(pq)}{n_{XY}}\right\}^{1/2}\right),$$

$$\left\|\hat{\mathbf{C}} - \mathbf{C}^*\right\|_1 \lesssim s_C\|\mathbf{C}^*\|_2^2\left(\frac{s_B\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{\min\left(n_{XX}^{1-\tau_1/2}, n_{XY}^{1-\tau_2/2}\right)} + \frac{\|\mathbf{B}^*\|_{L_1}^2(\log(pq))^{1/2}}{\min\left(n_{XX}^{1/2}, n_{XY}^{1/2}\right)}\right).$$

Next, we discuss some direct implications of Theorem 2. First, we show that our estimators are at least as good as the initial estimators under some conditions. Since $\tau_1, \tau_2 \le 1$ as $n_{jkl}^{XX/Y} \le n_{jk}^{XX}$ and $n_{jkl}^{XY/X} \le n_{jk}^{XY}$, the convergence rate of

$\|\hat{\mathbf{B}}-\mathbf{B}^*\|_F$ is no slower than $O_p(\max(\|\mathbf{B}^*\mathbf{C}^*\|_{L_1},1)\sqrt{s_B\log(pq)/\min(n_{XX},n_{XY})})$. Similarly, the convergence rate of $\|\hat{\mathbf{C}}-\mathbf{C}^*\|_F$ is no slower than $O_p(\sqrt{s_C}\|\mathbf{C}^*\|_2^2(\|\mathbf{B}^*\|_{L_1}^2+s_B\|\mathbf{B}^*\mathbf{C}^*\|_{L_1})\sqrt{\log(pq)/\min(n_{XX},n_{XY})})$. Here the two slowest convergence rates are achieved when $\tau_1=\tau_2=1$. If we assume $\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}=O(\|\mathbf{B}^*\|_{L_1}\sqrt{q})$, the upper bounds of $\|\hat{\mathbf{B}}-\mathbf{B}^*\|_F$ and $\|\hat{\mathbf{C}}-\mathbf{C}^*\|_F$ are at least as tight as $\|\hat{\mathbf{B}}_0-\mathbf{B}^*\|_F$ and $\|\hat{\mathbf{C}}_0-\mathbf{C}^*\|_F$.

On the other hand, if $\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}=o(\|\mathbf{B}^*\|_{L_1}\sqrt{q})$ or $\max(\tau_1,\tau_2)<1$ and $\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}^2=o(\min(n_{XX}^{1/2-\tau_1/2},n_{XY}^{1/2-\tau_2/2}))$, the upper bounds of $\|\hat{\mathbf{B}}-\mathbf{B}^*\|_F$ and $\|\hat{\mathbf{C}}-\mathbf{C}^*\|_F$ are strictly tighter than that of $\|\hat{\mathbf{B}}_0-\mathbf{B}^*\|_F$ and $\|\hat{\mathbf{C}}_0-\mathbf{C}^*\|_F$. One example is when $\mathrm{var}(\epsilon_j)>1/\sqrt{q}$ for all $j\le q$ and $cov(\epsilon_j,\epsilon_k)=0$ for $j\ne k$. Another example is when $n_{XX/Y}=o(n_{XX})$, $n_{XY/X}=o(n_{XY})$, and $\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}^2=o(\min(n_{XX}^{1/2-\tau_1/2},n_{XY}^{1/2-\tau_2/2}))$.

When $\mathbf{Y}$ is complete while $\mathbf{X}$ has missing entries, $\tau_1=-\infty$ and $\tau_2\in[0,1]$. Then convergence rate of $\hat{\mathbf{B}}$ in Theorem 2 becomes

$$\left\|\hat{\mathbf{B}}-\mathbf{B}^*\right\|_F\lesssim\sqrt{s_B}\left(\frac{\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{n_{XY}^{1-\tau_2/2}}+\left\{\frac{\log(pq)}{n_{XY}}\right\}^{1/2}\right).$$

When $\mathbf{X}$ are complete while $\mathbf{Y}$ have missing entries, $\tau_2=-\infty$ and $\tau_1\in[0,1]$. In this case, we can set $\alpha_1=\alpha_2=1$ and have

$$\left\|\hat{\mathbf{B}}-\mathbf{B}^*\right\|_F\lesssim\sqrt{s_B}\left(\frac{\|\mathbf{B}^*\mathbf{C}^*\|_{L_1}(\log(pq))^{1/2}}{n_{XX}^{1-\tau_1/2}}+\left\{\frac{\log(pq)}{n_{XY}}\right\}^{1/2}\right).$$

When both $\mathbf{X}$ and $\mathbf{Y}$ are complete, $\tau_1=\tau_2=-\infty$. In this case, we can set $\alpha_1=\alpha_2=\alpha_3=1$ and have

$$\|\hat{\mathbf{B}}-\mathbf{B}^*\|_F\lesssim\sqrt{\frac{s_B\log(pq)}{n}},\tag{3.6}$$

where n is the sample size. The error bound in (3.6) is the minimax rate of the ℓ_1-penalized estimator as shown in Raskutti, Wainwright and Yu (2011).

In Theorem 3 below, we show that our proposed method is model selection consistent.

Theorem 3. *Assume that Conditions* (A1)–(A5) *hold. Suppose* $1-\alpha_1=O(\sqrt{\log p/n_X}),1-\alpha_2=O(\sqrt{\log p/n_{XX}})$, $1-\alpha_3=O(\sqrt{\log(pq)/n_{XY}})$. *If* $(\log(pq)/n_{XY})^{1/2-\gamma_2}/\lambda_B=o(1),\lambda_B\|((\mathbf{C}^*\otimes\mathbf{\Sigma}_{XX})_{S_BS_B})^{-1}\|_{L_\infty}/\min_{j\in S_B}|\boldsymbol{\beta}_j^*|=o(1)$, $s_B\|((\mathbf{C}^*\otimes\mathbf{\Sigma}_{XX})_{S_BS_B})^{-1}\|_{L_\infty}(\log p/n_{XX})^{1/2-\gamma_2}=o(1)$, *and* $s_B(\log p/n_{XX})^{1/2-\gamma_1-\gamma_2}/\lambda_B=o(1)$, *then with probability at least* $1-4/p-4/(pq)-4/q$, *there exists a solution* $\hat{\mathbf{B}}$ *to* (2.5) *such that* $\mathrm{sign}(\hat{\mathbf{B}})=\mathrm{sign}(\mathbf{B}^*)$.

4. Numerical Study

In this section, we examine the performance of our proposed method (Multi-DISCOM) in terms of $\mathbf{\Sigma}_\epsilon$, the signal-to-noise ratio, and the distribution of the error $\boldsymbol{\epsilon}$ using numerical studies. We compare the efficiency of our proposed method with that of the following methods: (1) the complete lasso, which separately applies the lasso to each response using only samples with complete observations (both X and Y have no missing values); (2) the imputed lasso, which separately applies the lasso to each response using all samples, where missing data are imputed using the soft-thresholded SVD method; (3) the MBI, which separately applies the MBI (Xue and Qu (2021)) to each response using all samples, and the missing data are imputed using multiple block-wise imputation; (4) DISCOM, which separately applies the DISCOM method (Yu et al. (2020)) to each response; and (5) the imputed-MRCE, which runs the MRCE (Rothman, Levina and Zhu (2010)) using all samples, with missing data imputed using the soft-thresholded SVD method.

In all examples, we set $q = 4$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \sim N(\mathbf{0}, \mathbf{\Sigma})$, with $\sigma_{jt} = 0.6^{|j-t|}$. The ith row of the coefficient matrix $\mathbf{B}^*$ is $(1, 1.5, 1, 1.5)$, for $i = 1, p_1 + 1, p_1 + p_2 + 1$, and zero otherwise. The response $\mathbf{Y}$ has entries missing completely at random, with the missing proportion 0.01.

For each example, the data are generated from three modalities, with dimensions p_1, p_2, and p_3, respectively. The training data set contains n_1 samples with complete observations, n_2 samples from the third modality, n_3 samples from the first and third modalities, and n_4 samples from the first modality. The tuning data set contains 75 samples with complete observations, and the testing data set includes 300 samples with complete observations. For each method, we train our model with different tuning parameters on the training data set. Then we choose the optimal tuning parameter by minimizing the mean squared error (MSE) on the tuning data set.

For each example, we repeat the simulation 50 times. To evaluate the selection performance of the algorithm, we use the false-positive rate (FPR) and false-negative rate (FNR) as criteria: $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ and $\mathrm{FNR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP})$, where FN represents the number of coefficients wrongly detected as zero, TN is the number of coefficients correctly detected as zero, TP are is the number of coefficients correctly detected as nonzero and FP is the number of coefficients wrongly detected as nonzero. Furthermore, to evaluate the accuracy of our estimators, we use the MSE on the testing data set and the ℓ_2-distance $\|\hat{\mathbf{B}} - \mathbf{B}^*\|_F$ as criteria.

In Example 1, we examine our method related to $\mathbf{\Sigma}_\epsilon$. Let $n_1 = n_2 = n_3 = n_4 = 30$ and $p_1 = p_2 = p_3 = 30$. We set the error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq}) \sim N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$, with $\mathbf{\Sigma}_\epsilon = 3\mathbf{I}_2 \otimes \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$. We choose ρ between -0.4 and 0.4.

Table 1. Performance comparison for the methods in Example 1 with different ρ. The values in parentheses are the standard errors of the measures.

	Method	$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_F$	MSE	FPR	FNR
$\rho = -0.4$	Lasso	1.51(0.06)	3.70(0.06)	0.09(0.02)	**0.00(0.00)**
	Imputed-Lasso	1.73(0.06)	3.57(0.06)	0.11(0.01)	**0.00(0.00)**
	MBI	2.10(0.08)	4.26(0.09)	0.12(0.02)	0.11(0.03)
	DISCOM	1.44(0.04)	3.56(0.06)	0.05(0.00)	0.05(0.01)
	Imputed-MRCE	1.53(0.05)	3.72(0.08)	0.17(0.03)	0.08(0.02)
	Multi-DISCOM	**1.40(0.04)**	**3.39(0.08)**	**0.02(0.01)**	0.09(0.02)
$\rho = 0.4$	Lasso	1.55(0.06)	3.77(0.06)	0.11(0.02)	**0.00(0.00)**
	Imputed-Lasso	1.75(0.06)	3.61(0.06)	0.13(0.01)	**0.00(0.00)**
	MBI	2.14(0.08)	4.30(0.09)	0.13(0.02)	0.11(0.03)
	DISCOM	**1.46(0.04)**	3.59(0.06)	0.06(0.00)	0.05(0.01)
	Imputed-MRCE	1.54(0.05)	3.73(0.08)	0.19(0.03)	0.09(0.02)
	Multi-DISCOM	**1.43(0.04)**	**3.44(0.08)**	**0.04(0.01)**	0.07(0.02)

In Example 2, we examine the performance of our method related to the signal-to-noise ratio. Let $n_1 = n_2 = n_3 = n_4 = 30$ and $p_1 = p_2 = p_3 = 30$. We set the error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, with $\boldsymbol{\Sigma}_\epsilon = \alpha \mathbf{I}_2 \otimes \left(\begin{smallmatrix} 1 & -0.4 \\ -0.4 & 1 \end{smallmatrix}\right)$, and choose α between one and five.

In Example 3, we examine the robustness of our method when the error follows a heavy-tailed distribution. Let $n_1 = n_2 = n_3 = n_4 = 30$ and $p_1 = p_2 = p_3 = 30$. We set the error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq}) \sim t_{10}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, where $\boldsymbol{\Sigma}_\epsilon = 3\mathbf{I}_2 \otimes \left(\begin{smallmatrix} 1 & -0.4 \\ -0.4 & 1 \end{smallmatrix}\right)$, and $t_\nu(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ refers to Student's t distribution with location vector $\mathbf{0}$ and scale matrix $\boldsymbol{\Sigma}_\epsilon$.

To demonstrate the results, we focus on Example 1. We report the results of the other examples in the Supplementary Material.

The results in Table 1 indicate that the multi-DISCOM method delivers the best performance in all settings. Specifically, the multi-DISCOM method produces a smaller MSE and estimation errors than those of the other methods in all settings, especially when there are large correlations between the responses. In addition, the lasso method with imputed data may deliver worse selection performance, possibly because of the randomness in the imputation of the block-missing data. The results in Table 4 in the Supplementary Material indicate that the multi-DISCOM method has a greater advantage when the signal-to-noise ratio is small. When the ratio is smaller, the noise has a stronger effect on $\mathbf{Y}$, and hence considering the precision matrix is more more helpful for our estimation.

5. Application to the ADNI Study

We apply the multi-DISCOM method to data from the ADNI study (Mueller et al. (2005)), and compare it with several existing approaches. A primary goal of

Table 2. Performance comparison for the ADNI data.

Method	Overall MSE	MSE_{MMSE}	MSE_{ADAS1}	MSE_{ADAS2}	# of Selected Features
Lasso	93.37(3.82)	5.31(0.19)	29.84(1.35)	58.23(2.40)	54.20
Imputed-Lasso	80.40(1.62)	4.54(0.12)	25.80(0.51)	50.07(1.15)	165.00
MBI	91.84(3.02)	5.13(0.14)	28.43(1.17)	58.29(2.16)	59.87
DISCOM	67.47(1.33)	**4.26(0.11)**	21.76(0.51)	41.45(0.86)	72.87
Imputed-MRCE	67.41(2.02)	**4.29(0.10)**	**21.61(0.65)**	41.50(1.33)	218.50
Multi-DISCOM	**65.82(1.21)**	**4.22(0.12)**	**21.18(0.46)**	**40.41(0.80)**	89.67

this analysis is to identify biological markers and neuropsychological assessments to measure the progression of mild cognitive impairment (MCI) and early AD. We are interested in predicting the mini mental-state examination (MMSE), ADAS1, and ADAS2, which are common diagnotic scores for AD. The data processing steps are summarized in the Supplementary Material.

After data processing, we have 93 features from MRI, 93 features from PET, and five features from CSF. There are 805 subjects in total, including 199 subjects with complete MRI, PET, and CSF features, 197 subjects with MRI and PET features only, 201 subjects with MRI and CSF features only, and 208 subjects with MRI features only.

In our analysis, we divide the data into training, tuning, and testing sets. The training set consists of all subjects with incomplete observations and 40 randomly selected subjects with complete features. The tuning set consists of another 40 randomly selected subjects with complete observations. The testing set contains the remaining 119 subjects with complete observations. We train our model using different tuning parameters on the training set, choosing the tuning parameter that minimizes the MSE on the tuning set. The testing set is used to evaluate the methods. We use all methods shown in the simulation study to predict the MMSE score. For each method, the analysis is repeated 30 times using different partitions of the data. In addition to the sum of the MSE of all three responses, we compare the MSEs for each response (MSE_{MMSE}, MSE_{ADAS1}, and MSE_{ADAS2}) as criteria. We also compare the number of features selected by each method.

As shown in Table 2, the multi-DISCOM method outperforms all other methods. The DISCOM method has a similar overall MSE to that of the multi-DISCOM method, but worse MSE_{ADAS1} and MSE_{ADAS2}. One possible reason for this is that ADAS1 and ADAS2 are highly correlated, which means considering the precision matrix can help. Because there are 208 subjects with MRI features only, the MBI method may not impute those 208 subjects accurately. As a result, the MBI method may not perform well in this case.

With regard to model selection, both the DISCOM method and the multi-DISCOM method deliver relatively simple models. Figure 2 shows the selection frequency of the 191 features when predicting ADAS1. The selection frequency of

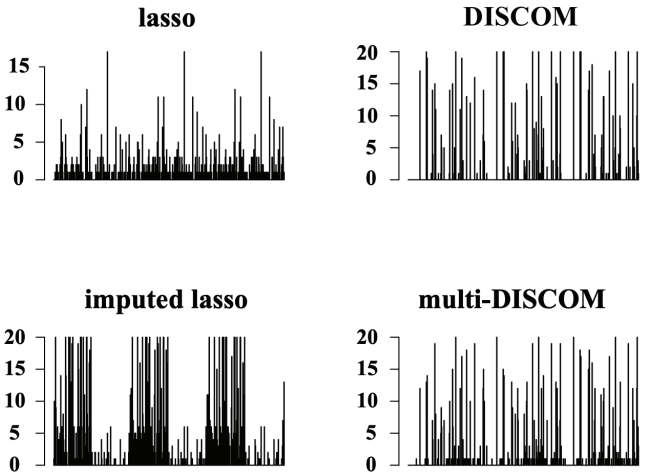

Figure 2. Selection frequency of 191 features for prediction of ADAS1 score.

each feature is defined as the number of times of it is selected in the 30 replications. As shown in Figure 2, for our method, some features are often selected, and many other features are rarely selected. Thus our method delivers robust model selection. However, the features selected by the imputed lasso method vary across replications. One possible reason for the unstable performance in terms of model selection is the randomness in the imputation of the block-missing data. Hippocampus formation left (69th region) and amygdale right (83th feature) are frequently selected by our method, and have been shown to be highly correlated with AD and MCI (Jack et al. (1999); Misra, Fan and Davatzikos (2009); Zhang and Shen (2012)); however, the DISCOM method rarely selects these features.

6. Conclusion

We have proposed a joint estimation method in a penalized framework with an entry-wise ℓ_1-regularization using block-missing multi-modal predictors. We first estimate the covariance matrix of the predictors using a linear combination of the estimates of the variance of each predictor, the estimates of the intra-modality covariance matrix, and the cross-modality covariance matrix. The proposed estimator of the covariance matrix can be positive semidefinite and more accurate than the sample covariance matrix. In the second step, we use the estimated covariance matrix and a penalized estimator to deliver a sparse estimate of the coefficients in the optimal linear prediction. We also establish the theory for the estimation and feature selection consistency. Extensive simulation studies indicate that our method exhibits promising performance in terms of estimation, prediction, and model selection for block-missing multi-modal data. Finally, we apply the multi-DISCOM method to the ADNI data set, showing that our model has good prediction power and meaningful interpretation.

Supplementary Material

The online Supplementary Material includes additional results of our numerical studies, technical conditions and proofs.

Acknowledgments

The authors thank the editor, associate editor, and reviewers for their helpful comments and suggestions. This research was supported in part by NSF grant DMS-2100729 and NIH grants R01GM126550 and R01AG073259. Haodong Wang gratefully acknowledges the partial support from the National Science Foundation, award NSF-DMS-1929298 to the Statistical and Applied Mathematical Sciences Institute.

References

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research* **9**, 485–516.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 3–54.

Cai, T. T., Li, H., Liu, W. and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.

Chen, J., Xu, P., Wang, L., Ma, J. and Gu, Q. (2018). Covariate adjusted precision matrix estimation via nonconvex optimization. In *Proceedings of the 35th International Conference on Machine Learning* 922–931.

Fisher, T. J. and Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics and Data Analysis* **55**, 1909–1918.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9**, 432–441.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* **5**, 248–264.

Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J. et al. (1999). Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* **52**, 1397–1397.

Johnson, C. R. (1990). Matrix completion problems: A survey. In *Matrix Theory and Applications*, 171–198. American Mathematical Society, Rhode Island.

Kim, S. and Xing, E. P. (2012). Tree-guided group Lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics* **6**, 1095–1117.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis* **111**, 241–255.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research* **16**, 559–616.

Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics* **7**, 495–522.

Misra, C., Fan, Y. and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD. *Neuroimage* **44**, 1415–1422.

Molstad, A. J., Sun, W. and Hsu, L. (2020). A covariance-enhanced approach to multi-tissue joint eQTL mapping with application to transcriptome-wide association studies. *arXiv:2001.08363*.

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

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57**, 6976–6994.

Rothman, A. J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.

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

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

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5**, 2630–2650.

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

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 329–346.

Zhang, D. and Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* **59**, 895–907.

Haodong Wang

Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260, USA.

E-mail: haodong@ad.unc.edu

Quefeng Li

Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260, USA.

E-mail: quefeng@email.unc.edu

Yufeng Liu

Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260, USA.

E-mail: yfliu@email.unc.edu

(Received May 2021; accepted July 2022)