| Statistica Sinica Preprint No: SS-2021-0170 | |
|---|---|
| **Title** | Multi-response Regression for Block-missing Multi-modal Data without Imputation |
| **Manuscript ID** | SS-2021-0170 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202021.0170 |
| **Complete List of Authors** | Haodong Wang, Quefeng Li and Yufeng Liu |
| **Corresponding Authors** | Yufeng Liu |
| **E-mails** | yfliu@email.unc.edu |
| Notice: Accepted version subject to English editing. | |

Statistica Sinica

# 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 paper, we consider the problem of parameter estimation and variable selection for multi-response regression using block-missing multi-modal data. Our method allows dimensions of both responses and predictors to be large, and the responses to be incomplete and correlated. Such a problem arises in many practical situations in the high-dimensional setting. Our proposed method includes two steps to make prediction from the multi-response linear regression model with block-missing multi-modal predictors. In the first step, without imputing the missing data, we make use of 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, based on the estimated covariance and cross-covariance matrices, we estimate both the precision matrix of the response vector given predictors and the sparse regression parameter matrix simultaneously by a penalized method. The effectiveness of the proposed method is demonstrated by theoretical studies, simulated examples, and the analysis of a multi-modal imaging dataset 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 has attracted growing research attentions in statistics and machine learning communities (Rothman et al., 2010; Lee and Liu, 2012; Loh and Zheng, 2013). While linear regression with a scalar response has been well studied, many applications may have a vector as the response. In particular, multi-response models have wide applications in scientific fields, especially for biological problems (Kim and Xing, 2012). For example, for multi-tissue joint expression quantitative trait loci (eQTL) mapping (Molstad et al., 2020), researchers consider predicting gene expression values in multiple tissues simultaneously by using a weighted sum of eQTL genotypes. Separate prediction for each tissue can be inefficient since same genes in different tissues are often correlated due to the shared genetic variants or other unmeasured common regulators. In order to use data from all tissues simultaneously, a joint eQTL modeling has been proposed to take cross-tissue expression dependence into account (Molstad et al., 2020).

To apply variable selection methods for multi-response problems, one could separately fit each response via a single-response model. There are many well-studied variable selection methods for the single-response linear regression model such as LASSO (Tibshirani, 1996). Although it is simple to apply a single-response linear regression method for each response separately, such a procedure neglects the dependency structure among responses. By incorporating the dependency structure of the response vector, one may obtain a more efficient multi-response linear regression approach in terms of estimation and prediction.

To handle multi-response regression problems, a well-known approach, the Curds and Whey, was proposed by Breiman and Friedman (1997) to improve the prediction performance by utilizing dependency among responses. Specifically, they first fit a single-response regression model for each response and then modify the predicted values from those regressions by shrinking them using canonical correlations between the response variables and the predictors. Another popular approach to handle multi-response regression is to use dimension reduction. In particular, reduced rank regression (Izenman, 1975) minimizes the least squares criterion subject to the constraint on the rank of regression parameter matrix. Yuan et al. (2007) further extended this method for the high dimensional setting. Their idea

is to obtain dimension reduction by encouraging sparsity among singular values of the parameter matrix. Although these methods may achieve better prediction performance than the separate univariate regression, they did not address the problem of variable selection.

In order to handle correlated responses together with variable selection, the precision matrix of response vector given predictors and the regression parameter matrix can be estimated separately or simultaneously (Lee and Liu, 2012). For separate estimation, Cai et al. (2013) used a constrained $\ell_1$ 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. One potential drawback of this indirect method is that it ignores the relationship between different responses given predictors when estimating the regression parameter matrix. In order to use all information more efficiently, it can be desirable to estimate the precision matrix and regression parameter matrix simultaneously. In the literature, various joint estimation techniques were studied by Rothman et al. (2010), Yin and Li (2011) and Lee and Liu (2012). They formulated the multi-response regression problem in a penalized log-likelihood framework, so that the parameter and precision matrices

can be estimated simultaneously. Using a similar idea, Chen et al. (2018) proposed an estimation procedure to estimate the parameter and precision matrices simultaneously based on the generalized Dantzig selector.

Despite a lot of development for multi-response linear regression, most existing methods only deal with complete data without missing entries. However, many practical data are incomplete, especially for multi-modal data. For instance, in the study of Alzheimer's Disease (AD), data from different sources are collected. This includes 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 due to patient dropouts or other practical issues. This leads to a block-wise missing data structure. It is important to integrate data from all modalities to improve model prediction and variable selection.

To handle incomplete multi-modal data, one may simply remove those observations with missing entries. However, such a procedure may greatly reduce the number of observations and lead to loss of information. Another approach is to perform data imputation. Existing imputation methods, such as matrix completion (Johnson, 1990) algorithms may possibly be unstable when the missing values happen in blocks. In order to deal with multi-modal block-wise missing data, Yu et al. (2020) proposed a new direct sparse re-

gression procedure using covariance from block-missing multi-modal data (DISCOM). They first used all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable. Based on the estimated covariance matrix and the estimated cross-covariance vector, they then used an extended Lasso-type estimator to estimate the coefficients. However, the DISCOM only considers single-response regression. Recently, Xue and Qu (2021) proposed the Multiple Block-wise Imputation (MBI) method for single-response regression when data are block-wise missing. They developed an estimating equation approach to accommodate block-wise missing patterns in multi-modal data. The method was shown to have high selection accuracy and low estimation error for single-response regression with block-wise missing data. However, since their imputation method requires analyzing all combinations of different blocks, it can be computationally expensive when the number of modalities is large.

In this paper, 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 responses and predictors and correlations among responses. This method can also handle the case that no subject has complete observations, while most traditional methods do not allow this. Our method

includes two steps. The first step is to estimate each element of the covariance and cross-covariance matrices by using all available observations without imputation. The second step is to use a penalized approach to estimate the sparse regression coefficient matrix and the precision matrix of the error terms simultaneously. We show that this method has estimation and model selection consistency under the high-dimensional setting. Numerical studies and the ADNI data application also 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. We present simulation studies and a multi-modal ADNI data example in Sections 4 and 5.

## 2. Methodology

### 2.1 Problem setup and notations

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 the $n \times q$ error matrix. We assume that $\{\mathbf{x}_i\}_{i=1}^n$ are 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 $k$-th 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 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 further 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 $S_C = \{j : \mathrm{vec}(\mathbf{C}^*)_j \neq 0\}$, where "vec" is the vectorization by 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 this article. The symbol $\mathbb{S}_+^{d\times d}$ is used to denote the 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 $\ell_1$–norm as $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$ and its entrywise $\ell_\infty$–

norm as $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$. In addition, we define its matrix $\ell_1$–norm as $\|\mathbf{A}\|_{L_1} = \max_j \sum_i |a_{ij}|$, matrix $\ell_\infty$–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 $a_n \leq c b_n$ 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

For the multi-response linear regression model (2.1), if one separately applies least squares estimation with the $\ell_1-$norm penalty to each response, it essentially solves

$$\arg\min_{\mathbf{B}} \mathbb{E}\left[\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2\right] + \lambda\|\mathbf{B}\|_1 = \arg\min_{\mathbf{B}} \operatorname{tr}\left(\tfrac{1}{2}\mathbf{B}^\top \mathbf{\Sigma}_{XX}\mathbf{B} - \mathbf{\Sigma}_{XY}^\top \mathbf{B}\right) + \lambda\|\mathbf{B}\|_1, \quad (2.2)$$

where $\lambda$ is a tuning parameter. We refer to this method as the separate LASSO, whose solution is denoted as $\hat{\mathbf{B}}^{LASSO}$. However, such an approach fails to account for the correlations between responses and may lead to poor

predictive performance (see, e.g., Breiman and Friedman (1997)). To produce a better estimator, we propose to incorporate $\mathbf{\Sigma}_\epsilon$ into the estimation of $\mathbf{B}^*$and solve the following problem:

$$\hat{\mathbf{B}}^0 = \arg\min_{\mathbf{B}} \operatorname{tr}\left[\mathbf{C}^*\hat{\mathbf{\Sigma}}_{\mathbf{YY}} + \mathbf{C}^*\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{\mathbf{XX}}\mathbf{B} - \mathbf{2}\mathbf{C}^*\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{\mathbf{XY}}\right] + \lambda\|\mathbf{B}\|_1, \quad (2.3)$$

where $\lambda$ is a tuning parameter, $\hat{\mathbf{\Sigma}}_{YY}$, $\hat{\mathbf{\Sigma}}_{XX}$ and $\hat{\mathbf{\Sigma}}_{XY}$ are some estimators of $\mathbf{\Sigma}_{YY}$, $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{XY}$.

In practice, $\mathbf{C}^*$ is also unknown. It is natural to estimate $\mathbf{C}^*$ first, then plug the estimate $\hat{\mathbf{C}}$ into (2.3) and solve the following problem:

$$\hat{\mathbf{B}}^0 = \arg\min_{\mathbf{B}} \operatorname{tr}\left[\hat{\mathbf{C}}\hat{\mathbf{\Sigma}}_{\mathbf{YY}} + \hat{\mathbf{C}}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{\mathbf{XX}}\mathbf{B} - \mathbf{2}\hat{\mathbf{C}}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{\mathbf{XY}}\right] + \lambda\|\mathbf{B}\|_1. \quad (2.4)$$

We refer this method as the two-step weighted LASSO. But as shown by the toy example in Section 2.2.1, the two-step weighted LASSO may perform worse than the separate LASSO in some problems.

In this article, we propose to estimate $\mathbf{B}^*$ and $\mathbf{C}^*$ simultaneously by solving the following optimization problem:

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

where $\lambda_B$ and $\lambda_C$ are tuning parameters. When $\lambda_C$ is large enough, Theorem 4 by Banerjee et al. (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 is no missing entries, our proposed method (2.5) reduces to the sparse conditional Gaussian graphical model introduced by Yin and Li (2011).

The toy example in Section 2.2.1 illustrates that our joint estimation model (2.5) has better estimation performance than the two-step weighted LASSO and the separate LASSO.

### 2.2.1 Toy example

For illustration, we consider a toy example similar to the one 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 $\rho$ is an unknown constant. We perform simulation studies for this example with 200 training samples, 300 tuning samples and 1000 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, the two-step weighted LASSO and the joint estimation model (2.5). In Case 1, the two-step weighted LASSO has a smaller estimation error than the separate LASSO when $\rho$ is positive. The result flips when $\rho$ is negative. While in Case 2, the separate LASSO has a smaller estimation error than the two-step weighted LASSO when $\rho$ is

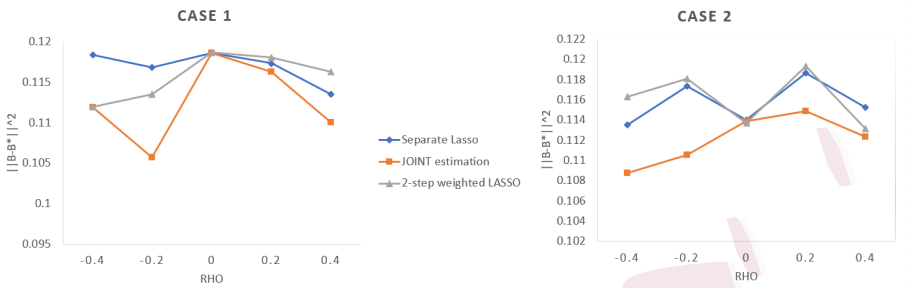


Figure 1: Plots of the estimation errors for separated LASSO, two-step weighted LASSO and joint estimation when $\mathbf{\Sigma}_\epsilon = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$. The left panel is for $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 3.5 \end{smallmatrix}\right)$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

positive. The joint estimation model performs the best in all cases.

The simulation results can be explained by the following calculations. With the penalty parameter $\lambda$, 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

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

When 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

$\text{sign}(\hat{B}^S_{i1}\hat{B}^S_{i2}) = -1$. Compared with $\hat{B}^{\text{LASSO}}_{ij} = \text{sign}(\hat{B}^S_{ij})[\hat{B}^S_{ij} - \lambda/2]_+$, $\hat{B}^{\text{2step}}_{ij}$ only differs in the shrinkage amount for each entry. The shrinkage amounts for all entries of the Separate LASSO are the same, which only depend on the tuning parameter $\lambda$. The shrinkage amounts for all entries of the two-step weighted LASSO depend on $\rho$, $\lambda$ 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 $\rho$ in Case 1, where $\mathbf{B}^* = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 3.5 \end{smallmatrix}\right)$. Since $B^*_{21}$, $B^*_{22}$ are far from 0, 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 $0.7\lambda$, while the shrinkage amounts for $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ are $0.3\lambda$. Thus the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are smaller than the shrinkage amounts for $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$. This means that with the tuning parameter $\lambda$ that shrinks $\hat{B}^{2step}_{11}$ and $\hat{B}^{2step}_{12}$ to 0, the shrinkage amounts for $\hat{B}^{2step}_{21}$ and $\hat{B}^{2step}_{22}$ are smaller than the shrinkage amounts for $\hat{B}^{LASSO}_{21}$ and $\hat{B}^{LASSO}_{22}$. Thus the two-step weighted LASSO has a smaller estimation error than 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 $0.3\lambda$, while the shrinkage amounts for $\hat{B}^{2step}_{11}$

and $\hat{B}_{12}^{2step}$ are $0.7\lambda$. This means that with the tuning parameter $\lambda$ that shrinks $\hat{B}_{11}^{2step}$ and $\hat{B}_{12}^{2step}$ to 0, the shrinkage amounts for $\hat{B}_{21}^{2step}$ and $\hat{B}_{22}^{2step}$ are larger than the shrinkage amounts for $\hat{B}_{21}^{LASSO}$ and $\hat{B}_{22}^{LASSO}$. Thus the separate LASSO is preferred to the two-step weighted LASSO in this scenario. When $\text{sign}(\hat{B}_{11}^{S}\hat{B}_{12}^{S}) = 1$, 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 separate LASSO only when $\rho$ is negative. In conclusion, the performance of the two-step weighted LASSO compared with 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) can perform better by choosing a large $\lambda_C$. Otherwise, the joint estimation model (2.5) can perform better by choosing a relatively small $\lambda_C$. Thus the joint estimation model (2.5) can perform competitively in all cases.

### 2.2.2 Covariance estimation

Next we introduce how to obtain $\hat{\mathbf{\Sigma}}_{XX}$, $\hat{\mathbf{\Sigma}}_{XY}$ and $\hat{\mathbf{\Sigma}}_{YY}$ when data have block-missing values. The following notation will be used in this article. For the $j$th predictor, define $S_j^X = \{i : x_{ij} \text{ is not missing}\}$. For the $j$th response,

define $S_j^Y = \{i : y_{ij} \text{ is not missing}\}$. Define $S_{jk}^{XX} = \{i : x_{ij} \text{ and } x_{ik} \text{ are not missing}\}$, $S_{jk}^{XY} = \{i : x_{ij} \text{ and } y_{ik} \text{ are not missing}\}$, $S_{jkl}^{XX/Y} = \{i : x_{ij}, x_{ik} \text{ are not missing but } y_{il} \text{ is missing}\}$, $S_{jkl}^{XY/X} = \{i : x_{ij}, y_{ik} \text{ are not missing but } x_{il} \text{ is missing}\}$ and $S_{jk}^{YY} = \{i : y_{ij} \text{ and } y_{ik} \text{ are not missing}\}$. Denote the cardinality of $S_j^X, S_j^Y$, $S_{jk}^{XX}$, $S_{jk}^{XY}$, $S_{jkl}^{XX/Y}$, $S_{jkl}^{XY/X}$ and $S_{jk}^{YY}$ as $n_j^X, n_j^Y$ $n_{jk}^{XX}$, $n_{jk}^{XY}$, $n_{jkl}^{XX/Y}$, $n_{jkl}^{XY/X}$, and $n_{jk}^{YY}$, respectively. Denote $n_X = \min_j |S_j^X|$, $n_{XX} = \min_{j,k} |S_{jk}^{XX}|$, $n_{XY} = \min_{j,k} |S_{jk}^{XY}|$, $n_{YY} = \min_{j,k} |S_{jk}^{YY}|$, $n_{XX/Y} = \max_{j,k,l} |S_{jkl}^{XX/Y}|$ and $n_{XY/X} = \max_{j,k,l} |S_{jkl}^{XY/X}|$.

We propose the initial estimators of $\mathbf{\Sigma}_{XX}$, $\mathbf{\Sigma}_{XY}$ and $\mathbf{\Sigma}_{YY}$ to be the sample covariance matrices using all available data, i.e. $\tilde{\mathbf{\Sigma}}_{XX} = (\tilde{\sigma}_{jt}^{XX}), \tilde{\mathbf{\Sigma}}_{XY} = (\tilde{\sigma}_{jt}^{XY}), \hat{\mathbf{\Sigma}}_{YY} = (\hat{\sigma}_{jt}^{YY})$, 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

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

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

For block-missing data $\mathbf{X}$, the above 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 $\tilde{\mathbf{\Sigma}}^{k_1k_2}$ for $1 \leq k_1, k_2 \leq K$ and $\tilde{\mathbf{\Sigma}}^{k_1k_2}$ being 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, which 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 the block-missing multi-modal data, due to the imbalanced sample sizes, 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 $\mathbf{\Sigma}_{XX}$ by a linear combination of $\tilde{\mathbf{\Sigma}}_I$ and $\tilde{\mathbf{\Sigma}}_C$ with different weights. In addition, to ensure positive definiteness of our estimation, we adopt the idea of shrinkage estimation of the covariance matrix (Fisher and

Sun, 2011) and add the diagonal matrix $\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)$ to our estimator,

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

where $\alpha_1, \alpha_2 \in [0,1]$ are two shrinkage weights. We add the diagonal matrix $\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)$ 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 $\alpha_1\lambda_{\min}(\tilde{\mathbf{\Sigma}}_I)+(1-\alpha_1)\lambda_{\min}(\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I))+\alpha_2\lambda_{\min}(\tilde{\mathbf{\Sigma}}_C)$. Since $\mathrm{diag}(\tilde{\mathbf{\Sigma}}_I)$ is a positive definite matrix, by carefully selecting the tuning parameters $\alpha_1$ and $\alpha_2$, the eigenvalues of our estimator can be guaranteed to be positive.

As we dicussed before, 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 to estimate $\mathbf{\Sigma}_{XY}$ by

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

where $\alpha_3 \in [0,1]$ is the shrinkage weight. We want to find the optimal linear combination $\hat{\mathbf{\Sigma}}^*_{XY} = \alpha^*_3 \tilde{\mathbf{\Sigma}}_{XY}$ whose expected quadratic loss $\mathbb{E}\|\hat{\mathbf{\Sigma}}^*_{XY} - \mathbf{\Sigma}_{XY}\|_F$ is minimized.

In our paper, we only consider a relative low dimension of $Y$ with not too many incomplete observations, so we will use $\hat{\mathbf{\Sigma}}_{YY}$ defined in (2.7) directly. But 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 $\theta = \|\operatorname{diag}(\tilde{\mathbf{\Sigma}}_I) - \mathbf{\Sigma}_I\|_F$. The optimal choice for the weights of $\alpha_1, \alpha_2$, and $\alpha_3$ is shown in the following proposition 2.1.

**Proposition 2.1.** *The solutions to the following two optimization problems:*

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

$$\alpha_3^* = \arg\min_{\alpha_3} \mathbb{E}\left\|\hat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY}\right\|_F^2 \tag{2.11}$$

*are*

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

*In addition, for* $\hat{\mathbf{\Sigma}}_{XX}^* = \alpha_1^* \tilde{\mathbf{\Sigma}}_I + (1 - \alpha_1^*) \operatorname{diag}(\tilde{\mathbf{\Sigma}}_I) + \alpha_2^* \tilde{\mathbf{\Sigma}}_C$ *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 $\ell_2$-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 2.1 shows that our estimator is more accurate than the sample covariance matrix.

Proposition 2.1 is closely related to Proposition 1 in Yu et al. (2020). They calculated the optimal weight and estimation error for their proposed estimator $\hat{\mathbf{\Sigma}}^*_{XX,DISCOM}$ of $\mathbf{\Sigma}_{XX}$, whose estimation error is

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

where $\tilde{\theta}^2 = \|\mathrm{tr}(\mathbf{\Sigma})\mathbf{I_p}/p - \mathbf{\Sigma}_I\|_F^2$. We can see that our estimator $\hat{\mathbf{\Sigma}}_{XX}$ has smaller $\ell_2$-error compared to their estimator. Comparing to their proposition, we also prove that 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 to solve the optimization problem (2.5). Since (2.5) is a bi-convex problem, the standard approach to solve this problem is via the alternating minimization method. In particular, starting with some given initial point $(\hat{\mathbf{B}}_0, \hat{\mathbf{C}}_0)$, at the $t$–th iteration, we solve solving the following problems

$$\hat{\mathbf{B}}_t = \arg\min_{\mathbf{B}} \mathrm{tr}\left[\hat{\mathbf{C}}_{t-1}\hat{\mathbf{\Sigma}}_{YY} + \hat{\mathbf{C}}_{t-1}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\hat{\mathbf{C}}_{t-1}\mathbf{B}^\top\hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_B\|\mathbf{B}\|_1, \tag{2.12}$$

$$\hat{\mathbf{C}}_t = \arg\min_{\mathbf{C}\in\mathbb{S}_+^{q\times q}} \mathrm{tr}\left[\mathbf{C}\hat{\mathbf{\Sigma}}_{YY} + \mathbf{C}\hat{\mathbf{B}}_{t-1}^\top\hat{\mathbf{\Sigma}}_{XX}\hat{\mathbf{B}}_{t-1} - 2\mathbf{C}\hat{\mathbf{B}}_{t-1}^\top\hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_C\|\mathbf{C}\|_1 - \log\det\mathbf{C}. \tag{2.13}$$

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). Since (2.12) is quadratic in $\mathbf{B}$, we use the

coordinate descent algorithm to solve it. Then we adopt the graphical lasso method by Friedman et al. (2008) to solve (2.13). We summarize the above procedures in Algorithm 1 below.

**Algorithm 1:** Alternating minimization updating algorithm

**Input:** $\mathbf{X}$, $\mathbf{Y}$, $\lambda_C, \lambda_B$

**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

$$\hat{\mathbf{B}}_0 = \arg\min_{\mathbf{B}} \operatorname{tr}\left[\hat{\mathbf{\Sigma}}_{YY} + \mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XY}\right] + \lambda_{B_0}\|\mathbf{B}\|_1, \quad (2.14)$$

$$\hat{\mathbf{C}}_0 = \arg\min_{\|\mathbf{C}\|_1 \le R, \mathbf{C}\in\mathbb{S}_+^{d\times d}} \operatorname{tr}(\mathbf{C}\hat{\mathbf{\Sigma}}_0) - \log\det(\mathbf{C}) + \lambda_{C_0}\|\mathbf{C}\|_1, \quad (2.15)$$

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

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

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

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

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

$$\lambda_C\|\mathbf{C}\|_1 - \log\det\mathbf{C},$$

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

## 3. Theoretical study

We establish the following theoretical results. First, we prove in Theorem 3.1 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 3.4. Finally, the selection consistency of our proposed method is shown in Theorem 3.5. 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 3.1, we prove the large deviation bounds for our proposed estimators $\hat{\mathbf{\Sigma}}_{XX}$, $\hat{\mathbf{\Sigma}}_{XY}$ and $\hat{\mathbf{\Sigma}}_{YY}$.

**Theorem 3.1.** *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 positive constants* $v_1'$, $v_2'$, *and* $v_3'$ *such that*

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\right\|_\infty \geq v_1'\sqrt{\frac{\log p}{n_{XX}}}\right) \leq \frac{4}{p}, \tag{3.16}$$

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY}\right\|_\infty \geq v_2'\sqrt{\frac{\log(pq)}{n_{XY}}}\right) \leq \frac{4}{pq}, \tag{3.17}$$

$$P\left(\left\|\hat{\mathbf{\Sigma}}_{YY} - \mathbf{\Sigma}_{YY}\right\|_\infty \geq v_3'\sqrt{\frac{\log q}{n_{YY}}}\right) \leq \frac{4}{q}. \tag{3.18}$$

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

$$\begin{aligned}
\left\|\tilde{\mathbf{\Sigma}}_{XX,\text{complete}} - \mathbf{\Sigma}_{XX}\right\|_\infty &= O_p\left(\sqrt{(\log p)/n_{\text{complete}}}\right), \\
\left\|\tilde{\mathbf{\Sigma}}_{XY,\text{complete}} - \mathbf{\Sigma}_{XY}\right\|_\infty &= O_p\left(\sqrt{(\log(pq))/n_{\text{complete}}}\right), \\
\left\|\tilde{\mathbf{\Sigma}}_{YY,\text{complete}} - \mathbf{\Sigma}_{YY}\right\|_\infty &= O_p\left(\sqrt{(\log q)/n_{\text{complete}}}\right),
\end{aligned}$$

where $n_{\text{complete}}$ is the number of samples with complete observations; see Yu et al. (2020). For block-missing data, $n_{\text{complete}}$ 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 3.2.** *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* $\lambda_{B_0} = C(\log(pq)/\min(n_{XY}, n_{XX}))^{\frac{1}{2}}\|\mathbf{B}^*\|_{L_1}$ *for some large enough constant* $C$, *then with probability at least* $1 - 4/p - 4/(pq)$, *the initial estimator*

$\hat{\mathbf{B}}_0 = \arg\min_{\mathbf{B}} \mathrm{tr}[\hat{\mathbf{\Sigma}}_{YY} + \mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XX}\mathbf{B} - 2\mathbf{B}^\top \hat{\mathbf{\Sigma}}_{XY}] + \lambda_B \|\mathbf{B}\|_1$ *satisfies*

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

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 $\|\mathbf{\Sigma}_{XX}^{-1}\|_{L_\infty}$. 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 $\|\mathbf{B}^*\|_{L_1} \asymp \|\mathbf{\Sigma}_{XX}^{-1}\|_{L_\infty}$, 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 3.2 as $n_{\text{complete}}$ 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 3.2 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 3.3.** *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*

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

There are two terms in the estimation error bound of $\hat{\mathbf{C}}_0$. The first term $\|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}^2\sqrt{\frac{s_C\log(pq)}{\min(n_{XX},n_{XY})}}$ comes from the error induced by using incomplete observations to estimate $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{XY}$. The second term $\|\mathbf{C}^*\|_2^2\|\mathbf{B}^*\|_{L_1}s_B\sqrt{\frac{s_C q\log(pq)}{\min(n_{XX},n_{XY})}}$ 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 3.4 below establishes the consistency of proposed estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ in (2.5).

**Theorem 3.4.** *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* $\lambda_B$ *and* $\lambda_C$ *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 3.4. 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 $O_p(\sqrt{s_C}\|\mathbf{C}^*\|_2^2(\|\mathbf{B}^*\|_{L_1}^2 + s_B\|\mathbf{B}^*\mathbf{C}^*\|_{L_1})\sqrt{\frac{\log(pq)}{\min(n_{XX}, n_{XY})}})$. 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 $\text{var}(\epsilon_j) > \frac{1}{\sqrt{q}}$ for all $j \leq q$ and $cov(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$. 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 3.4 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

$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_F \lesssim \sqrt{s_B \log(pq)/n}, \tag{3.19}$$

where $n$ is the sample size. The error bound in (3.19) is the minimax rate of the $\ell_1$-penalized estimator as shown in Raskutti et al. (2011).

In Theorem 3.5 below, we show that our proposed method is model selection consistent.

**Theorem 3.5.** *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* $(\log(pq)/n_{XY})^{\frac{1}{2}-\gamma_2}/\lambda_B = o(1)$, $\lambda_B\|((\mathbf{C}^*\otimes\mathbf{\Sigma}_{XX})_{S_BS_B})^{-1}\|_{L_\infty}/\min_{j\in S_B}|\boldsymbol{\beta}_j^*| = o(1)$, $s_B\|((\mathbf{C}^*\otimes\mathbf{\Sigma}_{XX})_{S_BS_B})^{-1}\|_{L_\infty}(\log p/n_{XX})^{\frac{1}{2}-\gamma_2} = o(1)$, *and* $s_B(\log p/n_{XX})^{\frac{1}{2}-\gamma_1-\gamma_2}/\lambda_B = o(1)$, *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) related to $\mathbf{\Sigma}_\epsilon$, the signal-to-noise ratio and the distribution of error $\boldsymbol{\epsilon}$ through some numerical studies. We compare the efficiency of our proposed method with some other methods. These methods include (1) Complete Lasso, which separately applies Lasso to each response only using samples with complete observations (both $X$ and $Y$ have no missing values); (2) Imputed-Lasso, which separately applies Lasso to each response using all samples, where missing data are imputed by the Soft-thresholded SVD method; (3) MBI, which separately applies the MBI (Xue and Qu, 2021) to each response using all samples, where missing data are imputed by the

Multiple Block-wise Imputation; (4) DISCOM, which separately applies the DISCOM (Yu et al., 2020) to each response; (5) Imputed-MRCE, which runs the MRCE (Rothman et al., 2010) using all samples with missing data imputed by the Soft-thresholded SVD method.

In all examples, we set $q = 4$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\sigma_{jt} = 0.6^{|j-t|}$. The $i$th 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 0 otherwise. The response $\mathbf{Y}$ has missing entries completely at random, with the missing proportion 0.01.

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

For each example, we repeat the simulation 50 times. To evaluate the selection performance of the algorithm, we use false-positive rate (FPR) and false-negative rate (FNR) as criteria: FPR = FP/(FP + TN) and FNR =

FN/(FN + TP), where FN represents the number of coefficients wrongly detected to be zero, TN are the number coefficients rightfully detected to be zero, TP are the coefficients rightfully detected to be nonzero and FP are the coefficients wrongly detected to be nonzero. Furthermore, to evaluate the accuracy of our estimators, we used the mean squared error (MSE) on the testing dataset and the $\ell_2$ distance $\|\hat{\mathbf{B}} - \mathbf{B}^*\|_F$ as criteria.

In Example 1, we examine our method related to $\mathbf{\Sigma}_\epsilon$. Let $n_1 = n_2 = n_3 = n_4 = 30$, $p_1 = p_2 = p_3 = 30$. We set 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 $\rho$ ranging from $-0.4$ to 0.4.

In Example 2, we examine the performance of our method related to the signal-to-noise ratio. Let $n_1 = n_2 = n_3 = n_4 = 30$, $p_1 = p_2 = p_3 = 30$. We set error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq}) \sim N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ with $\mathbf{\Sigma}_\epsilon = \alpha\mathbf{I}_2 \otimes \left(\begin{smallmatrix} 1 & -0.4 \\ -0.4 & 1 \end{smallmatrix}\right)$, and range $\alpha$ from 1 to 5.

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

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

The results in Table 1 indicate that the Multi-DISCOM delivers the best performance in all settings. Specifically, the Multi-DISCOM produces smaller MSE and estimation errors than the other methods in all settings, especially when the correlations between different responses are large. In addition, the Lasso method using the imputed data may deliver worse selection performance, possibly due to randomness involved in the imputation of block-missing data. The results in Table 4 in the Supplement Materials indicate that the Multi-DISCOM has more advantage when signal-to-noise ratio is small. When the signal-to-noise ratio is smaller, the noise has stronger effect on $\mathbf{Y}$ and hence taking the precision matrix into account is more helpful for our estimation.

## 5. Application to the ADNI study

We apply the Multi-DISCOM to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Mueller et al., 2005) and compare it with several existing approaches. A primary goal of this analysis is to identify biological markers and neuropsychological assessments to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). We are interested in predicting Mini-Mental State Examination (MMSE), ADAS1 and ADAS2. These scores are commonly used diagnotic scores of

| | 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) |

Table 1: Performance comparison of different methods for Example 1 with different $\rho$'s. The values in the parentheses are the standard errors of the measures.

AD. Data processing steps are summarized in the supplementary materials.

After data processing, we have 93 features from MRI, 93 features from PET and 5 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 with different tuning parameters on the training set. Then we choose the tuning parameter which minimizes the mean squared error on the tuning set. The testing set is used to evaluate different methods. We used all methods shown in the simulation study to predict the MMSE score. For each method, the analysis was repeated 30 times using different partitions of the data. In addition to the sum of mean squared errors (MSE) of all three responses, we compare MSEs for each response ($\text{MSE}_{MMSE}$, $\text{MSE}_{ADAS1}$ and $\text{MSE}_{ADAS2}$) as criteria. We also compare the number of features selected by each method.

| Method | Overall MSE | $\text{MSE}_{MMSE}$ | $\text{MSE}_{ADAS1}$ | $\text{MSE}_{ADAS2}$ | # 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 |

Table 2: Performance comparison for the ADNI data.

As shown in Table 2, the Multi-DISCOM delivers better performance than all other methods. The DISCOM has a similar overall MSE as the Multi-DISCOM, but worse $\text{MSE}_{ADAS1}$ and $\text{MSE}_{ADAS2}$. One possible reason is that ADAS1 and ADAS2 are highly correlated, so taking the precision matrix into account can help. Since there are 208 subjects with MRI features only, the MBI method may not impute those 208 subjects accurately. As a consequence, the MBI method may not perform well in this case.

Regarding to model selection, both the DISCOM and the Multi-DISCOM can deliver relatively simple models. Figure 2 shows the selection frequency of the 191 features when predicting ADAS1. The selection frequency of each feature is defined as the number of times of being selected in the 30 replications. As shown in Figure 2, for our method, some features are often

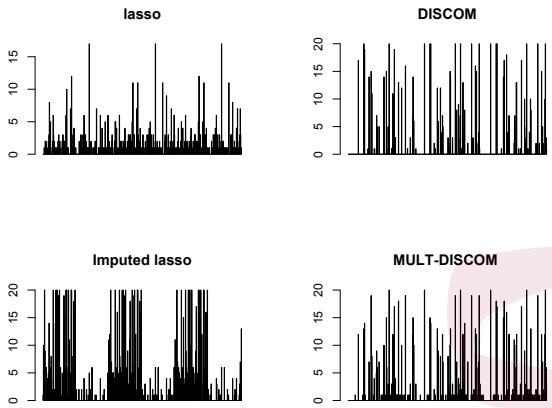


Figure 2: Selection frequency of 191 features for prediction of ADAS1 score.

selected and many other features are rarely selected. This means that our method could deliver robust model selection. However, for the Imputed-Lasso method, it selects very different features in different replications. One possible reason for the unstable performance on model selection is due to the randomness involved in the imputation of block-missing data. Hippocampus formation left (69th region) and amygdale right (83th feature) are frequently selected by our method and known to be highly correlated with AD and MCI by many existing studies (Jack et al., 1999; Misra et al., 2009; Zhang and Shen, 2012), but the DISCOM rarely selects these features.

## 6. Conclusion

In this paper, we propose a joint estimation method in a penalized framework with the entry-wise $\ell_1$ 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, based on the estimated covariance matrix, a penalized estimator is used to deliver a sparse estimate of the coefficients in the optimal linear prediction. Theoretical studies on the estimation and feature selection consistency are established. Extensive simulation studies also indicate that our method has promising performance on estimation, prediction and model selection for the block-missing multi-modal data. Finally, we apply the Multi-DISCOM to the ADNI dataset and demonstrate that our model has good prediction power and meaningful interpretation.

## Supplementary Materials

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

**Acknowledgments**

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

**References**

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (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 J. H. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(1), 3–54.

Cai, T. T., H. Li, W. Liu, and J. Xie (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika 100*(1), 139–156.

Chen, J., P. Xu, L. Wang, J. Ma, and Q. Gu (2018). Covariate adjusted

precision matrix estimation via nonconvex optimization. In *International Conference on Machine Learning*, pp. 922–931.

Fisher, T. J. and X. Sun (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics and Data Analysis 55*(5), 1909–1918.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis 5*(2), 248–264.

Jack, C. R., R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos, and E. Kokmen (1999). Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. *Neurology 52*(7), 1397–1397.

Johnson, C. R. (1990). Matrix completion problems: a survey. In *Matrix Theory and Applications*, Volume 40, pp. 171–198. Amer. Math. Soc.

Kim, S. and E. P. Xing (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics 6*(3), 1095–1117.

Lee, W. and Y. Liu (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 M. J. Wainwright (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research 16*(1), 559–616.

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

Misra, C., Y. Fan, and C. Davatzikos (2009). Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad. *Neuroimage 44*(4), 1415–1422.

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

Mueller, S. G., M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics 15*(4), 869–877.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over lq-balls. *IEEE Transactions on Information Theory 57*(10), 6976–6994.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

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

Xue, F. and A. Qu (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association 116*(536), 1914–1927.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*(4), 2630.

Yu, G., Q. Li, D. Shen, and Y. Liu (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association 115*(531), 1406–1419.

Yuan, M., A. Ekici, Z. Lu, and R. Monteiro (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(3), 329–346.

Zhang, D. and D. Shen (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage 59*(2), 895–907.

Haodong Wang

Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill

E-mail: (haodong@ad.unc.edu)

Quefeng Li

Department of Biostatistics, The University of North Carolina at Chapel Hill

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 Car-

olina at Chapel Hill

E-mail: (yfliu@email.unc.edu)