JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2019.1632079

Optimal Sparse Linear Prediction for Block-missing Multi-modality Data Without Imputation

Guan Yu[a], Quefeng Li[b], Dinggang Shen[c,d], and Yufeng Liu[e]

[a]Department of Biostatistics, State University of New York at Buffalo; [b]Department of Biostatistics, University of North Carolina at Chapel Hill; [c]Department of Radiology and BRIC, University of North Carolina at Chapel Hill; [d]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea; [e]Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Science, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC

ABSTRACT

In modern scientific research, data are often collected from multiple modalities. Since different modalities could provide complementary information, statistical prediction methods using multimodality data could deliver better prediction performance than using single modality data. However, one special challenge for using multimodality data is related to block-missing data. In practice, due to dropouts or the high cost of measures, the observations of a certain modality can be missing completely for some subjects. In this paper, we propose a new direct sparse regression procedure using covariance from multimodality data (DISCOM). Our proposed DISCOM method includes two steps to find the optimal linear prediction of a continuous response variable using block-missing multimodality predictors. In the first step, rather than deleting or imputing missing data, we make use of all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable. The proposed new estimate of the covariance matrix is a linear combination of the identity matrix, the estimates of the intra-modality covariance matrix and the cross-modality covariance matrix. Flexible estimates for both the sub-Gaussian and heavy-tailed cases are considered. In the second step, based on the estimated covariance matrix and the estimated cross-covariance vector, an extended Lasso-type estimator is used to deliver a sparse estimate of the coefficients in the optimal linear prediction. The number of samples that are effectively used by DISCOM is the minimum number of samples with available observations from two modalities, which can be much larger than the number of samples with complete observations from all modalities. The effectiveness of the proposed method is demonstrated by theoretical studies, simulated examples, and a real application from the Alzheimer's Disease Neuroimaging Initiative. The comparison between DISCOM and some existing methods also indicates the advantages of our proposed method.

ARTICLE HISTORY
Received April 2017
Revised December 2018

KEYWORDS
Block-missing; Huber's M-estimate; Lasso; Multi-modality; Prediction; Sparse regression.

1. Introduction

With the advance of modern scientific research, complex data are often collected from multiple modalities (sources or types). In neuroscience, different brain images such as magnetic resonance imaging (MRI) and positron emission tomography (PET) are used to study the brain structure and function. In biology, data from different modalities such as gene expressions and copy numbers are collected to understand the complex mechanism of cancers. Since different modalities could provide complementary information, statistical prediction methods using multimodality data could deliver better prediction performance than using single modality data. However, one special challenge for using multi-modality data is related to missing data, which is unavoidable due to some reasons such as the high cost of measures or the patients' dropout. Generally, the observations of a certain modality can be missing completely, that is, a complete block of the data is missing. One example of block-missing multi-modality data is shown in Figure 1. In this example, there are n samples (each row represents one sample), three modalities and one response variable. The blank regions with question mark indicate missing data. As shown in Figure 1, for many samples, the observations from some modality are missing completely. The number of samples with complete observations is much smaller than the sample size n.

To predict the response variable using the high-dimensional block-missing multimodality data, a common strategy is to use the Lasso (Tibshirani 1996) or some other penalized regression methods (e.g., Fan and Li 2001; Zou and Hastie 2005; Zhang 2010) only for the data with complete observations. However, this strategy can greatly reduce the sample size and waste a lot of useful information in the samples with missing data. Another strategy is to impute the missing data first by some existing imputation methods (Hastie et al. 1999; Cai et al. 2010). These methods can be effective when the positions of the missing data are random, but they can be unstable when a complete block of the data is missing. Recently, motivated by applications in genomic data integration, Cai et al. (2016) proposed a new

CONTACT Dinggang Shen ✉ dinggang_shen@med.unc.edu Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea; Yufeng Liu ✉ yfliu@email.unc.edu Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Science, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

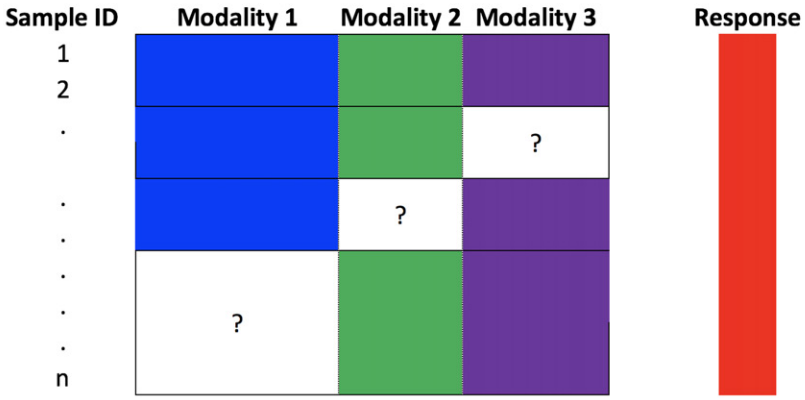

Figure 1. An illustration of block-missing multimodality data with three modalities.

framework of structured matrix completion to impute block-missing data. However, they only consider the case when the data are collected from two modalities. In the literature, rather than deleting or imputing missing data, some studies focus on using all available information. For example, Yuan et al. (2012) proposed the incomplete multisource feature learning (IMSF) method. The IMSF method performs regression on block-missing multimodality data without imputing missing data. It formulates the prediction problem as a multitask learning problem by first decomposing the prediction problem into a set of regression tasks, one for each combination of available modalities (e.g., modalities 1, 2, and 3; modalities 1 and 2; modalities 1 and 3; and modalities 2 and 3 for the example shown in Figure 1), and then building regression models for all tasks simultaneously. The important assumption in the IMSF method is that all models involving a specific modality share the common set of predictors for that particular modality. However, when different modalities are highly correlated, this assumption could be too strong. In that case, for some modalities, it can be more reasonable to choose different predictor subsets for different involved tasks. Therefore, it is desirable to develop flexible and efficient prediction methods applicable to block-missing multi-modality data.

In this article, we propose a new direct sparse regression procedure using covariance from multimodality data (DISCOM). For each sample, if some modality has missing entries, all the observations from that modality are missing simultaneously. Regardless of the underlying true model, we aim to find the optimal linear prediction for the response variable using the block-missing multi-modality data without imputing the missing data. Our method includes two steps. In the first step, we use all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable. The proposed new estimate of the covariance matrix is a linear combination of the identity matrix, the estimates of the intra-modality covariance matrix and the cross-modality covariance matrix. Flexible estimates for both the sub-Gaussian and heavy-tailed cases are considered. Many existing high-dimensional covariance estimation methods such as Bickel and Levina (2008), Cai and Liu (2011), Rothman (2012), Lounici et al. (2014), and Cai and Zhang (2016) can be used in this step. In the second step, based on the estimated covariance matrix and the estimated cross-covariance vector, we use an extended Lasso-type estimator to estimate the coefficients in the optimal linear prediction.

Note that there are some existing sparse regression methods in the literature using the estimation of the covariance matrix. For example, Jeng and Daye (2011) proposed the covariance-thresholded Lasso for complete data to improve variable selection by using the sparsity of the covariance matrix. Loh and Wainwright (2012) and Datta et al. (2017) proposed new estimators for the high dimensional regression with corrupted predictors, where all entries of the design matrix are assumed to be noisy or missing randomly and independently. The missing data problem they considered can be viewed as a special case of the block-missing multimodality data where each modality has only one predictor. To the best of our knowledge, there are no existing methods using a similar idea to DISCOM tailored for high-dimensional block-missing multimodality data. To investigate DISCOM, we have carefully studied its theoretical and numerical performance. For both the sub-Gaussian and heavy-tailed cases, we establish the consistency of estimation and model selection for the optimal linear predictor regardless of the underlying true model. Our theoretical studies indicate that DISCOM could make use of all available information of the block-missing multi-modality data effectively. The number of samples that are effectively used by DISCOM is the minimum number of samples with available observations from two modalities, which can be much larger than the number of samples with complete observations from all modalities. The comparison between DISCOM and some existing methods using simulated data and Alzheimer's Disease Neuroimaging Initiative (ADNI) data (www.loni.ucla.edu/ADNI) further demonstrate the effectiveness of our proposed method.

The rest of this article is organized as follows. In Section 2, we motivate and introduce our method. In Section 3, we show some theoretical results about the estimates of the covariance matrix, the cross-covariance vector and the coefficients in the optimal linear prediction for both the sub-Gaussian and heavy-tailed cases. The results about the model selection consistency are also provided. In Sections 4 and 5, we demonstrate the performance of our method on the simulated data and the ADNI dataset. We conclude this article in Section 6 and provide all technical proofs in the appendix.

2. Motivation and Methodology

We first show the motivation and the outline of our proposed method in Section 2.1. In Section 2.2, we introduce the proposed estimate of the covariance matrix of the predictors, and the estimate of the cross-covariance vector between the predictors and the response variable using the block-missing multi-modality data. In Section 2.3, we introduce Huber's M-estimate for the heavy-tailed case. In Section 2.4, we provide the estimation procedure for the coefficients in the optimal linear prediction.

The following notation will be used in this article. For a matrix $\mathbf{A} \in R^{m\times n}$, we use $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_{\max}$, and $\|\mathbf{A}\|_\infty$ to denote the Frobenius norm $\sqrt{\sum_{ij} a_{ij}^2}$, the max norm $\max_{ij} |a_{ij}|$, and the infinity norm $\max_i \sum_{j=1}^n |a_{ij}|$, respectively. For a vector $b \in R^{m\times 1}$, we use $\|b\|_2$, $\|b\|_{\max}$, and $\|b\|_1$ to denote the ℓ_2 norm $\sqrt{\sum_i b_i^2}$, the max norm $\max_i |b_i|$, and the ℓ_1 norm $\sum_{i=1}^n |b_i|$, respectively. In addition, we use $\text{sign}(\cdot)$ to denote the function that maps a positive entry to 1, a negative entry to -1, and 0 to 0.

2.1. Motivation

Suppose the predictors are collected from K modalities. For $k \in \{1, 2, \ldots, K\}$, there are p_k predictors from the k-th modality. Let n denote the sample size, $Y = (y_1, y_2, \ldots, y_n)^T$ denote the $n \times 1$ response vector centered to have mean 0, and $\mathbf{X}^{(k)} \in R^{n\times p_k}$ denote the design matrix of the p_k predictors from the kth modality. In addition, let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}) = (x_1, x_2, \ldots, x_n)^T$ denote the $n \times p$ design matrix, where $p = p_1 + p_2 + \cdots + p_K$. We assume that x_i's are iid. generated from some multivariate distribution with mean $0_{p\times 1}$ and covariance matrix Σ. We use $C = \text{cov}(x_i, y_i) = (c_1, c_2, \ldots, c_p)^T \in R^p$ to denote the cross-covariance vector between x_i and y_i.

To predict the response variable y using all predictors $X_1, X_2, \ldots, X_p$, we consider the optimal linear predictor $\hat{y} = \sum_{j=1}^p X_j\beta_j^0$, where the coefficient vector

$$\beta^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_p^0)^T$$
$$= \arg\min_\beta E\left[\left(y - \sum_{j=1}^p X_j\beta_j\right)^2\right] = \Sigma^{-1}C. \qquad (1)$$

The above coefficient vector β^0 can be viewed as the solution to the following optimization problem:

$$\min_\beta \frac{1}{2}\beta^T\Sigma\beta - C^T\beta.$$

If we know the true covariance matrix Σ and the true cross-covariance vector C, and assume that β^0 is sparse, we can estimate β^0 by solving the following optimization problem:

$$\min_\beta \frac{1}{2}\beta^T\Sigma\beta - C^T\beta + \lambda\|\beta\|_1, \qquad (2)$$

where λ is a nonnegative tuning parameter.

Motivated by (2), for the high-dimensional block-missing multimodality data, we propose a new method with two steps. In the first step, we use all available observations to estimate the covariance matrix Σ and the cross-covariance vector C. The estimates of Σ and C are denoted as $\hat{\Sigma}$ and $\hat{C}$, respectively. This step is very important to make full use of the block-missing multimodality data. In the second step, we estimate β^0 by solving the following optimization problem:

$$\min_\beta \frac{1}{2}\beta^T\hat{\Sigma}\beta - \hat{C}^T\beta + \lambda\|\beta\|_1. \qquad (3)$$

2.2. Standard Estimates of Σ and C

Considering block-missing multimodality data, for each sample, if a certain modality has missing entries, all the observations from that modality are missing. For each predictor j, define $S_j = \{i : x_{ij} \text{ is not missing}\}$. For each pair of predictors j and t, define $S_{jt} = \{i : x_{ij} \text{ and } x_{it} \text{ are not missing}\}$. The number of elements in S_j and S_{jt} are denoted as n_j and n_{jt}, respectively.

For the missing data mechanism, we only need to assume that for each predictor, the first sample moment and the second sample moment using all available observations are unbiased estimators of the first theoretical moment and the second theoretical moment of the distribution, respectively. This assumption is satisfied if we assume that each modality is missing completely at random. However, different predictors in the same modality are missing simultaneously. Under this assumption, for each $j \in \{1, 2, \ldots, p\}$, the available observations of the jth predictor are centered to have mean 0. A natural initial unbiased estimate of Σ using all available data is the sample covariance matrix

$$\tilde{\Sigma} = (\tilde{\sigma}_{jt})_{j,t=1,2,\ldots,p}, \text{ where } \tilde{\sigma}_{jt} = \frac{1}{n_{jt}}\sum_{i\in S_{jt}} x_{ij}x_{it}.$$

For the block-missing multi-modality data, the above initial estimate $\tilde{\Sigma}$ may have negative eigenvalues due to the unequal sample sizes n_{jt}'s. Therefore, it is not a good estimate of the covariance matrix Σ and not suitable to be used in (3) directly. It is important to find an estimator that is both positive semidefinite and more accurate than the initial estimate $\tilde{\Sigma}$.

According to the partition of the predictors into K modalities, the initial estimate of the covariance matrix $\tilde{\Sigma}$ can be partitioned into K^2 blocks, denoted by $\tilde{\Sigma}_{jt}$'s, where $j, t \in \{1, 2, \ldots, p\}$ and $\tilde{\Sigma}_{jt}$ is a $p_j \times p_t$ matrix. We denote

$$\tilde{\Sigma}_I = \begin{pmatrix} \tilde{\Sigma}_{11} & & & \\ & \tilde{\Sigma}_{22} & & \\ & & \ddots & \\ & & & \tilde{\Sigma}_{KK} \end{pmatrix},$$

$$\tilde{\Sigma}_C = \begin{pmatrix} 0 & \tilde{\Sigma}_{12} & \ldots & \tilde{\Sigma}_{1K} \\ \tilde{\Sigma}_{21} & 0 & \ldots & \tilde{\Sigma}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\Sigma}_{K1} & \tilde{\Sigma}_{K2} & \ldots & 0 \end{pmatrix},$$

where $\tilde{\Sigma}_I$ is called the intra-modality sample covariance matrix which is a $p \times p$ block-diagonal matrix containing K main diagonal blocks of $\tilde{\Sigma}$, and $\tilde{\Sigma}_C = \tilde{\Sigma} - \tilde{\Sigma}_I$ is called the cross-modality sample covariance matrix containing all the off-diagonal blocks of $\tilde{\Sigma}$. We also let Σ_I and Σ_C denote the true intra-modality covariance matrix and cross-modality covariance matrix, respectively. As shown in Figure 1, since the observations of some modalities are missing completely for many samples, there are more available samples to estimate the intra-modality covariance matrix Σ_I than the cross-modality covariance matrix Σ_C. Intuitively, it is relatively easier to estimate Σ_I than Σ_C. In view of this characteristic of the block-missing multi-modality data and the possible negative eigenvalues of $\tilde{\Sigma}$, we propose to use the following estimator:

$$\hat{\Sigma} = \alpha_1 \tilde{\Sigma}_I + \alpha_2 \tilde{\Sigma}_C + \alpha_3 \mathbf{I_p},$$

where α_1, α_2 and α_3 are three nonrandom weights, and $\mathbf{I_p}$ is a $p \times p$ identity matrix. Considering all possible linear combinations, we can find the optimal linear combination $\tilde{\Sigma}^* = \alpha_1^* \tilde{\Sigma}_I + \alpha_2^* \tilde{\Sigma}_C + \alpha_3^* \mathbf{I_p}$ whose expected quadratic loss $E[\| \tilde{\Sigma}^* - \Sigma \|_F^2]$ is the minimum. The optimal weights α_1^*, α_2^* and α_3^* are shown in the following Proposition 1. As a remark, Proposition 1 and all the theoretical analysis in Section 3 are conditional on the given missing pattern of different modalities.

Proposition 1. Consider the following optimization problem:

$$\min_{\alpha_1,\alpha_2,\alpha_3} E[\| \hat{\Sigma} - \Sigma \|_F^2] \qquad \text{subject to} \qquad \hat{\Sigma} = \alpha_1 \tilde{\Sigma}_I + \alpha_2 \tilde{\Sigma}_C + \alpha_3 \mathbf{I_p},$$

where the weights α_1, α_2 and α_3 are nonrandom. Denote $\gamma^* = \mathrm{tr}(\Sigma)/p$, $\delta_I^2 = E[\| \tilde{\Sigma}_I - \Sigma_I \|_F^2]$, $\delta_C^2 = E[\| \tilde{\Sigma}_C - \Sigma_C \|_F^2]$, and $\theta^2 = \| \gamma^* \mathbf{I_p} - \Sigma_I \|_F^2$. The optimal weights are

$$\alpha_1^* = \frac{\theta^2}{\theta^2 + \delta_I^2} \in [0,1], \qquad \alpha_2^* = \frac{\| \Sigma_C \|_F^2}{\| \Sigma_C \|_F^2 + \delta_C^2} \in [0,1],$$
$$\alpha_3^* = \gamma^*(1 - \alpha_1^*).$$

In addition, we have

$$\begin{aligned} E[\| \tilde{\Sigma}^* - \Sigma \|_F^2] &= \frac{\delta_I^2 \theta^2}{\delta_I^2 + \theta^2} + \frac{\delta_C^2 \| \Sigma_C \|_F^2}{\delta_C^2 + \| \Sigma_C \|_F^2} \leq \delta_I^2 + \delta_C^2 \\ &= E[\| \tilde{\Sigma} - \Sigma \|_F^2]. \end{aligned}$$

Proposition 1 shows that $\tilde{\Sigma}^*$ is more accurate than $\tilde{\Sigma}$. The relative improvement in the expected quadratic loss over the sample covariance matrix is equal to

$$\begin{aligned} \frac{E[\| \tilde{\Sigma} - \Sigma \|_F^2] - E[\| \tilde{\Sigma}^* - \Sigma \|_F^2]}{E[\| \tilde{\Sigma} - \Sigma \|_F^2]} &= \frac{\delta_I^2}{\delta_I^2 + \delta_C^2} \cdot (1 - \alpha_1^*) \\ &\quad + \frac{\delta_C^2}{\delta_I^2 + \delta_C^2} \cdot (1 - \alpha_2^*). \end{aligned}$$

Therefore, if $\tilde{\Sigma}_I$ is relatively accurate (δ_I^2 is small), the optimal weight $\alpha_1^* = \frac{\theta^2}{\theta^2 + \delta_I^2}$ should be large and the percentage of the relative improvement tends to be small. We can also make the same conclusions about $\tilde{\Sigma}_C$. For the block-missing multi-modality data, due to the unequal sample sizes, the initial estimate $\tilde{\Sigma}_I$ can be relatively accurate while the estimate $\tilde{\Sigma}_C$ is relatively inaccurate. It's reasonable to use different weights for $\tilde{\Sigma}_I$ and $\tilde{\Sigma}_C$. As a remark, Proposition 1 can be viewed as a generalization of Theorem 2.1 shown in Ledoit and Wolf (2004), where they studied the optimal linear combination of the sample covariance matrix and the identity matrix to estimate the covariance matrix for the complete data.

Regarding the cross-covariance vector C, we can use the following estimate

$$\tilde{C} = (\tilde{c}_1, \tilde{c}_2, \cdots, \tilde{c}_p)^T, \text{ where } \tilde{c}_j = \frac{1}{n_j} \sum_{i \in S_j} y_i x_{ij}.$$

Note that we use all available information to estimate Σ and C. The theoretical properties of $\tilde{\Sigma}$ and $\tilde{C}$ will be discussed in Section 3.

2.3. Robust Estimates of Σ and C

When the predictors and the response variable follow a sub-Gaussian distribution with an exponential tail, $\tilde{\Sigma}^*$ and $\tilde{C}$ introduced in Section 2.2 generally perform well. However, when the distributions of the predictors and the response variable are heavy-tailed, $\tilde{\Sigma}^*$ and $\tilde{C}$ may have poor performance, and therefore some robust estimates of Σ and C are required.

In this section, we introduce robust estimates of Σ and C based on Huber's M-estimator (Huber (1964)). In general, suppose $Z_1, Z_2, \ldots, Z_n$ are iid copies of a random variable Z with mean μ. Huber's M-estimator of μ is defined as the solution to the following equation:

$$\sum_{i=1}^{n} \psi_H(Z_i - \mu) = 0,$$

where $\psi_H(\cdot)$ is the Huber function which is given by

$$\psi_H(z) = \begin{cases} z & \text{if } |z| \leq H, \\ H \cdot \operatorname{sign}(z) & \text{otherwise.} \end{cases}$$

Using Huber's M-estimator, for the block-missing multimodality data, we can construct a robust initial estimate of Σ denoted by

$$\begin{aligned} \breve{\Sigma} &= (\breve{\sigma}_{jt})_{j,t=1,2,\ldots,p}, \text{ where } \breve{\sigma}_{jt} \\ &= \text{the solution to } \sum_{i \in S_{jt}} \psi_{H_{jt}}(x_{ij} x_{it} - \mu) = 0. \end{aligned}$$

In general, the parameters H_{jt} used in the Huber function can be chosen to be 1.345 in order to guarantee 95% efficiency relative to the sample mean if the data-generating distribution is Gaussian (Huber (1964)). However, for the block-missing multi-modality data, considering different numbers of samples available to estimate different entries of Σ, we propose to use different values of H flexibly. The choice of H_{jt} will be discussed in Section 3. Based on the robust initial estimate $\breve{\Sigma}$, we can use a similar idea introduced in Section 2.2 to find the optimal linear combination $\breve{\Sigma}^* = \alpha_1^* \breve{\Sigma}_I + \alpha_2^* \breve{\Sigma}_C + \alpha_3^* \mathbf{I_p}$ whose expected quadratic loss $E[\| \breve{\Sigma}^* - \Sigma \|_F^2]$ is the minimum. Similarly, we can use Huber's M-estimator to deliver a robust estimate of C which is defined as

$$\begin{aligned} \breve{C} &= (\breve{c}_1, \breve{c}_2, \cdots, \breve{c}_p)^T, \text{ where } \breve{c}_j \\ &= \text{the solution to } \sum_{i \in S_j} \psi_{H_j}(x_{ij} y_i - \mu) = 0. \end{aligned}$$

Here, we also propose to use different values of H when estimating different c_j's. The choice of H_j will be discussed in Section 3. The theoretical properties of $\breve{\Sigma}$ and $\breve{C}$ will be also shown in that section.

2.4. Estimate of β^0 in the Optimal Linear Prediction

After getting an initial estimate of Σ and C, for example, $\tilde{\Sigma}$ and $\tilde{C}$ (or $\breve{\Sigma}$ and $\breve{C}$), our proposed DISCOM method estimates β^0 by solving the following optimization problem:

$$\min_{\beta} \frac{1}{2}\beta^T \left[\alpha_1 \tilde{\Sigma}_I + \alpha_2 \tilde{\Sigma}_C + (1-\alpha_1)\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p} \right] \beta - \tilde{C}^T\beta + \lambda\|\beta\|_1, \tag{4}$$

where $\alpha_1 \in [0,1], \alpha_2 \in [0,1]$ are two weights and $\mathrm{tr}(\tilde{\Sigma})/p$ is used to estimate γ^*. In practice, both $\alpha_1 \in [0,1]$, $\alpha_2 \in [0,1]$, and λ can be chosen by cross-validation or an additional tuning dataset. To guarantee that the estimated covariance matrix $\hat{\Sigma} = \alpha_1 \tilde{\Sigma}_I + \alpha_2 \tilde{\Sigma}_C + (1-\alpha_1)\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p}$ is positive semidefinite, we need to choose reasonable α_1 and α_2 from the set $\{(\alpha_1,\alpha_2) : \alpha_1 \in [0,1], \alpha_2 \in [0,1]$, and $\lambda_{\min}(\hat{\Sigma}) \geq 0\}$, where $\lambda_{\min}(\hat{\Sigma})$ is the smallest eigenvalue of $\hat{\Sigma}$.

Besides the above tuning parameter selection method that searches for the best values of three parameters, we can use an efficient tuning method incorporating our theoretical results in Section 3. Our theoretical studies show that the tuning parameters α_1 and α_2 should satisfy the conditions $1-\alpha_1 = O(\sqrt{(\log p)/\min_j n_j})$ and $1-\alpha_2 = O(\sqrt{(\log p)/\min_{j,t} n_{jt}})$, respectively. Denote $m_1 = \sqrt{(\log p)/\min_j n_j}$ and $m_2 = \sqrt{(\log p)/\min_{j,t} n_{jt}}$. We can choose $\alpha_1 = 1-k_0 m_1$ and $\alpha_2 = 1-k_0 m_2$, where $k_0 \in [k_{\min}, k_{\max}]$ is a tuning parameter. To guarantee that both α_1 and α_2 are nonnegative, we set $k_{\max} = \min\{1/m_1, 1/m_2\}$. In addition, a reasonable value of k_0 should satisfy the following two conditions: (1) $\alpha_1 = 1-k_0 m_1 \leq 1$ and $\alpha_2 = 1-k_0 m_2 \leq 1$; (2) the estimate of the covariance matrix $\hat{\Sigma}$ is positive semidefinite. The first condition requires that $k_0 \geq 0$. If the smallest eigenvalue of the initial estimate $\tilde{\Sigma}$, denoted by $\lambda_{\min}(\tilde{\Sigma})$, is nonnegative, we can show that $\hat{\Sigma}$ is positive semidefinite for any nonnegative k_0. If $\lambda_{\min}(\tilde{\Sigma}) < 0$, since the smallest eigenvalue of $\hat{\Sigma}$ satisfies

$$\lambda_{\min}(\hat{\Sigma}) \geq \lambda_{\min}(\tilde{\Sigma}) + k_0 \cdot \left[\lambda_{\min}\left((m_2-m_1)\tilde{\Sigma}_I + m_1\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p} \right) - m_2\lambda_{\min}(\tilde{\Sigma}) \right],$$

to guarantee that $\hat{\Sigma}$ is positive semidefinite, we only need to require that

$$k_0 \geq -\lambda_{\min}(\tilde{\Sigma}) \Big/ \left[-m_2 \cdot \lambda_{\min}(\tilde{\Sigma}) + \lambda_{\min}\left((m_2-m_1)\tilde{\Sigma}_I + m_1\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p} \right) \right].$$

Therefore, if $\lambda_{\min}(\tilde{\Sigma}) \geq 0$, we choose $k_{\min} = 0$. Otherwise, we choose

$$k_{\min} = -\lambda_{\min}(\tilde{\Sigma}) \Big/ \left[-m_2 \cdot \lambda_{\min}(\tilde{\Sigma}) + \lambda_{\min}\left((m_2-m_1)\tilde{\Sigma}_I + m_1\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p} \right) \right].$$

For the block-missing multimodality data, since $m_2 \geq m_1 > 0$, we know that the matrix $(m_2-m_1)\tilde{\Sigma}_I + m_1\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p}$ is positive definite and therefore $k_{\min}$ is always less than $k_{\max} = \min\{1/m_1, 1/m_2\} = 1/m_2$.

By choosing $\alpha_1 = 1-k_0 m_1$ and $\alpha_2 = 1-k_0 m_2$, our proposed fast tuning parameter selection method searches the best value of $k_0 \in [k_{\min}, k_{\max}]$ and the parameter λ rather than searching three parameters α_1, α_2 and λ. In addition, instead of using the eigendecomposition for each parameter combination to check whether $\hat{\Sigma}$ is positive semidefinite, this method only requires two eigendecompositions of the matrices $\tilde{\Sigma}$ and $(m_2-m_1)\tilde{\Sigma}_I + m_1\frac{\mathrm{tr}(\tilde{\Sigma})}{p}\mathbf{I_p}$ before the tuning parameter selection process. For each $k_0 \in [k_{\min}, k_{\max}]$, we can incorporate the coordinate descent algorithm (Friedman et al. 2010) on a grid of λ values, from the largest one down to the smallest one, using warm starts. Alternatively, since $\hat{\Sigma}$ is positive semidefinite, we can use the LARS algorithm shown in Jeng and Daye (2011) to compute the solution path.

As many existing high-dimensional linear regression studies for the random design, we use the assumption $E(X) = 0$ to make our presentation more convenient. Our proposed DISCOM method can be used for the general case where $E(X) \neq 0$. In that case, we first center the available observations of each predictor and use $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p$ to denote the sample means of those p predictors. We also center the observed responses and use $\bar{Y}$ to denote the sample mean of the response variable. Let $\hat{\beta}$ denote the estimated regression coefficient vector calculated from the centered data. Our final predictive model is $\bar{Y} + \sum_{j=1}^{p}(X_j^* - \bar{X}_j)\hat{\beta}_j$, where $(X_1^*, X_2^*, \ldots, X_p^*)$ is a test data point. In practice, if our data are collected at various time points by different laboratories using multiple platforms, the iid assumption may be violated due to batch-effects. In that case, we suggest to use some existing statistical methods (e.g., the exploBATCH R package) to diagnose, quantify and correct batch effects before using our proposed DISCOM method.

3. Theoretical Study

Without loss of generalization, we assume that the true variances of all predictors, $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}$, are equal to 1 in our theoretical studies. For each $j \in \{1, 2, \ldots, p\}$, we assume that the observations of the predictor j are scaled such that $\sum_{i\in S_j} x_{ij}^2 = n_j$. In that case, we have $\tilde{\sigma}_{jj} = 1$. For Huber's M-estimator $\breve{\Sigma}$, we redefine $\breve{\sigma}_{jj}$ to be 1 for each j. Let $\tilde{\beta}$ and $\breve{\beta}$ denote the solutions to (4) using the sample covariance and Huber's M-estimator, respectively. We assume that β^0 is sparse and denote $J = \{j : \beta_j^0 \neq 0\}$ as the index set of the important predictors. Denote $s = |J|$ as the number of important predictors. Let $\beta_{\max}^0 = \max_{j\in J}|\beta_j^0|$ and $\beta_{\min}^0 = \min_{j\in J}|\beta_j^0|$. In Sections 3.1 and 3.2, we will discuss the theoretical properties in the sub-Gaussian case and the heavy-tailed case, respectively. The model selection consistency of our proposed method will be shown in Section 3.3.

3.1. Sub-Gaussian Case

The following conditions are considered in this section:

(A1) Suppose that there exists a constant $L > 0$ such that

$$E(\exp(tX_j)) \le \exp\left(\frac{L^2t^2}{2}\right) \text{ for all } j \in \{1,2,\ldots,p\} \text{ and } t \in R,$$

$$E(\exp(ty)) \le \exp\left(\frac{L^2t^2}{2}\right) \text{ for all } t \in R.$$

(A2) Suppose that the true covariance matrix Σ satisfies the following restricted eigenvalue (RE) condition:

$$\min_{\delta\in\{u\in R^p:\|u_{J^c}\|_1\le 7\|u_J\|_1\}} \frac{\delta^T\Sigma\delta}{\delta^T\delta} \ge m > 0.$$

Under condition (A1), the predictors and the response variable follow sub-Gaussian distributions with exponentially bounded tails. In this case, we propose to use $\tilde{\Sigma}$ and $\tilde{C}$ shown in Section 2.2 as the initial estimate of the covariance matrix Σ and the cross-covariance vector C, respectively. The RE condition (**A2**) is often used to obtain bounds of statistical error of the Lasso estimate (Datta et al. 2017). The following Theorem 1 shows the large deviation bounds of $\tilde{\Sigma}$ and $\tilde{C}$.

Theorem 1. Under condition (A1), if $\min_{j,t} n_{jt} \ge 6\log p$, there exists two positive constants $\nu_1 = 8\sqrt{6}(1+4L^2)$ and $\nu_2 = 4$ such that

$$\max_{j,t} P\left(|\tilde{\sigma}_{jt} - \sigma_{jt}| \ge \nu_1\sqrt{\frac{\log p}{n_{jt}}}\right) \le \frac{\nu_2}{p^3},$$

$$P\left(\|\tilde{\Sigma} - \Sigma\|_{\max} \ge \nu_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}\right) \le \frac{\nu_2}{p}.$$

There exists another two positive constants $\nu_3 = 16(1 + 4\frac{L^2}{\min\{\mathrm{var}(y),1\}})\max\{\mathrm{var}(y),1\}$ and $\nu_4 = 4$ such that

$$\max_{j} P\left(|\tilde{c}_j - c_j| \ge \nu_3\sqrt{\frac{\log p}{n_j}}\right) \le \frac{\nu_4}{p^2},$$

$$P\left(\|\tilde{C} - C\|_{\max} \ge \nu_3\sqrt{\frac{\log p}{\min_j n_j}}\right) \le \frac{\nu_4}{p}.$$

Remark 1. In our theoretical studies, we assume that the dimension p goes to infinity as the sample size $\min_{j,t} n_{jt}$ increases. If we further assume that $(\log p)/\min_{j,t} n_{jt} = o(1)$, the condition $\min_{j,t} n_{jt} > 6\log p$ is satisfied if the sample size $\min_{j,t} n_{jt}$ is sufficiently large. Then, Theorem 1 shows that $\|\tilde{\Sigma} - \Sigma\|_{\max} = O_p(\sqrt{(\log p)/\min_{j,t} n_{jt}})$. The performance of $\tilde{\Sigma}$ depends on the worst case when there are only $\min_{j,t} n_{jt}$ samples to estimate some entries in Σ. In addition, the convergence rate of $\|\tilde{C} - C\|_{\max}$ is $O_p(\sqrt{(\log p)/\min_j n_j})$. The performance of $\tilde{C}$ also depends on the worst case when there are only $\min_j n_j$ samples to estimate the covariance between some predictor and the response variable. Furthermore, if we only use samples with complete observations, using a similar proof, we can show that $\|\tilde{\Sigma} - \Sigma\|_{\max} = O_p(\sqrt{(\log p)/n_{\text{complete}}})$ and $\|\tilde{C} - C\|_{\max} = O_p(\sqrt{(\log p)/n_{\text{complete}}})$, where n_{complete} is the number of samples with complete observations. For the block-missing multimodality data, since n_{complete} can be much smaller than $\min_{j,t} n_{jt}$ and $\min_j n_j$, Theorem 1 indicates that the first step of our proposed DISCOM method can make full use of all available information. Based on the results shown in Theorem 1, we will show the convergence rate of $\|\tilde{\beta} - \beta^0\|_2$.

Theorem 2. Under conditions (**A1**) and (**A2**), let $1 - \alpha_1 = O(\sqrt{(\log p)/\min_j n_j})$ and $1 - \alpha_2 = O(\sqrt{(\log p)/\min_{j,t} n_{jt}})$. If $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$ and we choose $\lambda = 2\|\tilde{C} - \hat{\Sigma}\beta^0\|_{\max}$, then we have $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\lambda) = O_p(\|\beta^0\|_1 \sqrt{s(\log p)/\min_{j,t} n_{jt}})$.

Remark 2. As shown in the above Theorem 2, we have $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\|\tilde{C} - \hat{\Sigma}\beta^0\|_{\max})$. If we assume that (a) there is no missing data, (b) the predictors are generated from a multivariate Gaussian distribution, and (c) the true model is $Y = \mathbf{X}\beta^0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$. Then we will use $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ and $\tilde{C} = \mathbf{X}^TY/n$ to estimate Σ and C, respectively. Therefore, we have $\|\tilde{C} - \hat{\Sigma}\beta^0\|_{\max} = \|\mathbf{X}^T\boldsymbol{\epsilon}/n\|_{\max} = O_p(\sqrt{(\log p)/n})$, and $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{(s\log p)/n})$, which is the minimax ℓ_2-norm rate as shown in Raskutti et al. (2011). Since the complete data generated from the Gaussian random design can be viewed as a special type of block-missing multimodality data, the error bound in Theorem 2 is sharp.

On the other hand, if the true relationship between the conditional expectation $E(y|X_1, X_2, \ldots, X_p)$ and the predictors is nonlinear, we have $\tilde{C} - \hat{\Sigma}\beta^0 \neq \mathbf{X}^T\boldsymbol{\epsilon}/n$ and $\|\tilde{C} - \hat{\Sigma}\beta^0\|_{\max} = O_p(\|\beta^0\|_1\sqrt{(\log p)/n})$ as shown in the proof. In this case, if we still use the Lasso method to estimate the regression coefficients β^0 in the optimal linear predictor, we have $\|\tilde{\beta}_{\text{Lasso}} - \beta^0\|_2 = O_p(\|\beta^0\|_1\sqrt{s(\log p)/n})$. For the blocking missing multimodality data, since the Lasso method can only use the data with complete observations, we have $\|\tilde{\beta}_{\text{Lasso}} - \beta^0\|_2 = O_p(\|\beta^0\|_1\sqrt{s(\log p)/n_{\text{complete}}})$. However, as shown in Theorem 2, for our proposed DISCOM estimate $\tilde{\beta}$, we have $\|\tilde{\beta} - \beta^0\|_2 = O_p(\|\beta^0\|_1\sqrt{s(\log p)/\min_{j,t} n_{jt}})$. In practice, the minimum number of samples with available observations from two modalities ($\min_{j,t} n_{jt}$) can be much larger than the number of samples with complete observations from all modalities (n_{complete}). Theorem 2 indicates that DISCOM could make use of the block-missing multimodality data more effectively than the Lasso method using only the complete data.

In Theorem 2, the assumption $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$ is used to guarantee that $\hat{\Sigma}$ satisfies the RE condition with a high probability if the true covariance matrix Σ satisfies the RE condition (**A2**). Note that many existing sparse linear regression studies focus on the fixed design where the design matrix $\mathbf{X}$ is considered to be fixed and complete. In that case, $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ is assumed to satisfy the RE condition directly. For the general random design, Van De Geer and Bühlmann (2009) showed that $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ satisfies the RE condition as long as the true covariance matrix Σ satisfies the RE condition and $s^2\log p/n = o(1)$. For the special Gaussian random design, by a global analysis of the full random matrix $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ rather than a local analysis looking at individual entries of $\hat{\Sigma}$, Raskutti et al. (2010) shows that the matrix $\hat{\Sigma}$ satisfies the RE condition with a high probability if the true covariance matrix of the multivariate Gaussian distribution satisfies the RE condition and $n >$ Constant $\cdot\, s\log p$. In our article, since we consider the general random design including both sub-Gaussian distributions and heavy-tailed distributions, and study the proposed estimated covariance matrix $\hat{\Sigma}$ where $\hat{\Sigma} \neq \mathbf{X}^T\mathbf{X}/n$ in most cases, we use the condition $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$ to guarantee that the RE condition is satisfied with a high probability. This condition is very similar to the condition $s^2 \log p/n = o(1)$ used in Van De Geer and Bühlmann (2009) for the complete data.

For the general random design and the block-missing multimodality data, it is difficult to develop a weak condition (e.g., $s \log p/\min_{j,t} n_{jt} = o(1)$) using a similar global analysis of the full random matrix $\hat{\Sigma}$ as shown in Raskutti et al. (2010). Instead of using the condition $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$, we can use the following weak condition

$$\min_{j,t} n_{jt} > (128\nu_1'/m)^2(s^2 \log p),$$

where $\nu_1' > \nu_1$ is a positive constant. This condition is also used in some existing studies about random designs (Bühlmann and van der Geer 2011; Zhou et al. 2009).

3.2. Heavy-Tailed Case

In this section, we consider the heavy-tailed case. Instead of assuming that the distributions of the predictors and the response variable have exponential tails, we consider the following moment condition.

(A3) Suppose that $\max_{1\leq j\leq p} \mathrm{E}(X_j^4) \leq Q_1^2/48$ and $\mathrm{E}(y^4) \leq Q_2^2$, where Q_1 and Q_2 are two positive constants.

Condition (A3) assumes that the fourth moments of all predictors X_j's and the response variable y are bounded. Under condition (A3), the tails of the distributions of X_j's and y may not be exponentially bounded. In the literature on Lasso, most studies consider the fixed design (Meinshausen and Bühlmann 2006; Zhao and Yu 2006; Zou 2006) and the noise is usually assumed to be Gaussian (Meinshausen and Bühlmann 2006; Zhang and Huang 2008), or admits exponentially bounded tail (Bunea 2008; Meinshausen and Yu 2009). In this study, we consider a random design case and relax the distribution of X_j's and y to have finite fourth moments.

Next, we discuss the theoretical properties of Huber's M-estimators $\breve{\Sigma}$ and $\breve{C}$. Based on the convergence rates of $\|\breve{\Sigma} - \Sigma\|_{\max}$ and $\|\breve{C} - C\|_{\max}$, we will show the convergence rate of $\|\breve{\beta} - \beta^0\|_2$.

Theorem 3. Under condition (A3), let $H_{jt} = \frac{Q_1}{12}\sqrt{n_{jt}/\log p}$ for each $j, t \in \{1, 2, \ldots, p\}$, if $\min_{j,t} n_{jt} \geq 24 \log p$, we have

$$\max_{j,t} P\left(|\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{n_{jt}}}\right) \leq \frac{2}{p^3},$$

$$P\left(\|\breve{\Sigma} - \Sigma\|_{\max} \geq Q_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}\right) \leq \frac{2}{p}.$$

In addition, let $H_j = (Q_1 + Q_2)\sqrt{n_j/\log p}$ for each $j \in \{1, 2, \ldots, p\}$, we have

$$\max_{j} P\left(|\breve{c}_j - c_j| \geq 8(Q_1 + Q_2)\sqrt{\frac{\log p}{n_j}}\right) \leq \frac{2}{p^2},$$

$$P\left(\|\breve{C} - C\|_{\max} \geq 8(Q_1 + Q_2)\sqrt{\frac{\log p}{\min_j n_j}}\right) \leq \frac{2}{p}.$$

Remark 3. If we assume that $(\log p)/\min_{j,t} n_{jt} = o(1)$, the condition $\min_{j,t} n_{jt} > 24 \log p$ is satisfied if the sample size $\min_{j,t} n_{jt}$ is sufficiently large. Therefore, we have $\|\breve{\Sigma} - \Sigma\|_{\max} = O_p(\sqrt{(\log p)/\min_{j,t} n_{jt}})$ and $\|\breve{C} - C\|_{\max} = O_p(\sqrt{(\log p)/\min_j n_j})$. This indicates that Huber's M-estimators for the heavy-tailed case acquire the same convergence rate as the sample covariance estimates for the sub-Gaussian case. However, as shown in the next theorem, if the distributions of the predictors X_j's and the response variable y are not assumed to have exponentially bounded tails, the large deviation bounds of $\tilde{\Sigma}$ and $\tilde{C}$ can be wider than the bounds of Huber's M-estimators $\breve{\Sigma}$ and $\breve{C}$, respectively.

Theorem 4. Suppose $\max_{1\leq j\leq p} \mathrm{E}(X_j^{4\ell}) \leq T$ and $\mathrm{E}(y^{4\ell}) \leq T$, where $T > 0$, $\ell > 1$ are two constants. Then we have

$$\max_{j,t} P\left(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \frac{d_1}{2T}\sqrt{\frac{p}{n_{jt}}}\right) \leq \frac{d_2}{p^{2h}},$$

$$P\left(\|\tilde{\Sigma} - \Sigma\|_{\max} \geq \frac{d_1}{2T}\sqrt{\frac{p}{\min_{j,t} n_{jt}}}\right) \leq \frac{d_2}{p^{2h-2}},$$

where $d_1 > 0$, $d_2 > 0$, $h \in (1, \ell)$ are some constants. Furthermore,

$$\max_{j} P\left(|\tilde{c}_j - c_j| \geq \frac{d_3}{2T}\sqrt{\frac{p}{n_j}}\right) \leq \frac{d_4}{p^{2h-1}},$$

$$P\left(\|\tilde{C} - C\|_{\max} \geq \frac{d_3}{2T}\sqrt{\frac{p}{\min_j n_j}}\right) \leq \frac{d_4}{p^{2h-2}},$$

where $d_3 > 0$ and $d_4 > 0$ are two constants.

Remark 4. Under the moment condition, Theorem 4 shows that $\|\tilde{\Sigma} - \Sigma\|_{\max} = O_p(\sqrt{p/\min_{j,t} n_{jt}})$ and $\|\tilde{C} - C\|_{\max} = O_p(\sqrt{p/\min_j n_j})$. According to Proposition 6.2 in Catoni (2012), the bounds shown in Theorem 4 are actually tight. If the dimension p is very large, the large deviation bounds of $\|\tilde{\Sigma} - \Sigma\|_{\max}$ and $\|\tilde{C} - C\|_{\max}$ can be much larger than the bounds of $\|\breve{\Sigma} - \Sigma\|_{\max}$ and $\|\breve{C} - C\|_{\max}$, respectively. This necessitates the usage of a robust estimator.

In the next theorem, based on the large deviation bounds of $\|\breve{\Sigma} - \Sigma\|_{\max}$ and $\|\breve{C} - C\|_{\max}$, we show the convergence rate of $\|\breve{\beta} - \beta^0\|_2$.

Theorem 5. Under conditions (A2) and (A3), let $1 - \alpha_1 = O(\sqrt{(\log p)/\min_j n_j})$, $1 - \alpha_2 = O(\sqrt{(\log p)/\min_{j,t} n_{jt}})$, $H_{jt} = \frac{Q_1}{12}\sqrt{n_{jt}/\log p}$ and $H_j = (Q_1 + Q_2)\sqrt{n_j/\log p}$. If $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$ and let $\lambda = 2\|\breve{C} - \hat{\Sigma}\beta^0\|_{\max}$, then we have $\|\breve{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\lambda) = O_p(\|\beta^0\|_1 \sqrt{s(\log p)/\min_{j,t} n_{jt}})$.

Remark 5. Instead of using the condition $s\sqrt{(\log p)/\min_{j,t} n_{jt}} = o(1)$, we can assume that

$$\min_{j,t} n_{jt} > (128Q_1'/m)^2(s^2 \log p),$$

where $Q_1' > Q_1$ is a positive constant. Theorem 5 indicates that for the heavy-tailed case, under (**A3**), the convergence rate of

$\|\breve{\beta}-\beta^0\|_2$ is also $O_p(\|\beta^0\|_1\sqrt{(s\log p)/\min_{j,t} n_{jt}})$, which is the same as the rate shown in Theorem 2 under the sub-Gaussian assumption. However, as shown in our simulation study, if the response variable and the predictors follow sub-Gaussian distributions, DISCOM using standard estimates $\tilde{\Sigma}$ and $\tilde{C}$ generally has better finite sample performance than the method using robust estimates $\breve{\Sigma}$ and $\breve{C}$.

Remark 6. If we assume that p is fixed, for the sub-Gaussian case considered in Section 3.1, we can show that $\|\tilde{\Sigma}-\Sigma\|_{\max} = O_p((\min_{j,t} n_{jt})^{-1/2})$ and $\|\tilde{C} - C\|_{\max} = O_p((\min_j n_j)^{-1/2})$ according to Lemma 1 in Ravikumar et al. (2011) and a very similar proof of Theorem 1. For the heavy-tailed case considered in Section 3.2, if we assume that p is fixed, we can also show that $\|\breve{\Sigma} - \Sigma\|_{\max} = O_p((\min_{j,t} n_{jt})^{-1/2})$ and $\|\breve{C} - C\|_{\max} = O_p((\min_j n_j)^{-1/2})$ according to Theorem 5 in Fan et al. (2017) and a very similar proof of Theorem 3. Then, using the same proof of Theorem 2, we can also show that $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\lambda) = O_p(\sqrt{s}\|\tilde{C} - \hat{\Sigma}\beta^0\|_{\max})$. Since $\|\tilde{\Sigma} - \Sigma\|_{\max} = O_p((\min_{j,t} n_{jt})^{-1/2})$, $\|\tilde{C} - C\|_{\max} = O_p((\min_j n_j)^{-1/2})$, and p is fixed, we can further show that $\|\tilde{\beta} - \beta^0\|_2 = O_p(\beta^0_{\max}(\min_{j,t} n_{jt})^{-1/2})$. Similarly, for the heavy-tailed case, we can also show that $\|\breve{\beta} - \beta^0\|_2 = O_p(\beta^0_{\max}(\min_{j,t} n_{jt})^{-1/2})$. Therefore, the convergence rate of the estimation error in the classical fixed p setting is faster than the rate in the high dimensional setting where p grows to infinity.

3.3. Model Selection Consistency

In this section, we show that our proposed DISCOM method is model selection consistent. The following condition is considered.

(A4) $\|\Sigma_{J^cJ}\Sigma_{JJ}^{-1}\|_\infty \le 1-\eta$, where $\eta \in (0,1)$ is a constant, Σ_{J^cJ} is the sub-matrix of Σ with row indices in the set J^c and column indices in the set J, and Σ_{JJ} is the sub-matrix of Σ with both row and column indices in the set J.

Condition (A4) can be viewed as a population version of the strong irrepresentable condition proposed in Zhao and Yu (2006). In the following Theorem 6 and Theorem 7, we will show that our proposed DISCOM method is model selection consistent for the sub-Gaussian case and the heavy-tailed case, respectively.

Theorem 6. Under Conditions (A1) and (A4), let $1-\alpha_1 = O(\sqrt{(\log p)/\min_j n_j})$ and $1-\alpha_2 = O(\sqrt{(\log p)/\min_{j,t} n_{jt}})$. If $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{\frac{s^2\log p}{\min_{j,t} n_{jt}}} \longrightarrow 0$, and

$$\frac{1+s\beta^0_{\max}}{\lambda}\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}} \longrightarrow 0, \qquad \frac{\lambda\cdot\|(\Sigma_{JJ})^{-1}\|_\infty}{\beta^0_{\min}} \longrightarrow 0,$$

then there exists a solution $\tilde{\beta}$ to (4) such that $P(\mathrm{sign}(\tilde{\beta}) = \mathrm{sign}(\beta^0)) \longrightarrow 1$, as $\min_{jt} n_{jt} \to \infty$ and $p \to \infty$.

Remark 7. Note that the condition $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{(s^2\log p)/\min_{j,t} n_{jt}} = o(1)$ is used to guarantee that (a) $\|(\hat{\Sigma}_{JJ})^{-1}\|_\infty \le \text{Constant}\cdot\|(\Sigma_{JJ})^{-1}\|_\infty$ and (b) $\|\hat{\Sigma}_{J^cJ}\hat{\Sigma}_{JJ}^{-1}\|_\infty \le 1-\eta'$ if $\|\Sigma_{J^cJ}\Sigma_{JJ}^{-1}\|_\infty \le 1-\eta$ for the general random design with a high probability, where $\eta' \in (0,1)$ and $\eta \in (0,1)$ are two constants. For the fixed design, we do not need this condition. For the special Gaussian random design, as shown in Wainwright (2009), using some concentration inequalities about the normal distribution and the fact that $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ for the complete data, we can obtain model selection consistency with $n > \text{Constant}\cdot s\log(p-s)$. In our theoretical studies, since we consider the general random design including both sub-Gaussian distributions and heavy-tailed distributions, and $\hat{\Sigma} \ne \mathbf{X}^T\mathbf{X}/n$ for the block-missing multi-modality data, we use the condition $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{(s^2\log p)/\min_{j,t} n_{jt}} = o(1)$ to guarantee that (a) and (b) are satisfied. Note that this condition was also used in some existing model selection consistency studies for random designs (Jeng and Daye 2011; Datta et al. 2017).

As shown in the proof of Theorem 6, to guarantee that (a) and (b) are satisfied, instead of requiring $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{(s^2\log p)/\min_{j,t} n_{jt}} = o(1)$, we can use the following weak condition

$$\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{\frac{s^2\log p}{\min_{j,t} n_{jt}}} \le \frac{\eta}{\nu_1'(4+\eta)},$$

where $\nu_1' > \nu_1$ is a positive constant.

Theorem 7. Under conditions (A3) and (A4), let $H_{jt} = \frac{Q_1}{12}\sqrt{n_{jt}/\log p}$, $H_j = (Q_1+Q_2)\sqrt{n_j/\log p}$, $1-\alpha_1 = O(\sqrt{(\log p)/\min_j n_j})$, $1-\alpha_2 = O(\sqrt{(\log p)/\min_{j,t} n_{jt}})$. If $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{(s^2\log p)/\min_{j,t} n_{jt}} \longrightarrow 0$, and

$$\frac{1+s\beta^0_{\max}}{\lambda}\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}} \longrightarrow 0, \qquad \frac{\lambda\cdot\|(\Sigma_{JJ})^{-1}\|_\infty}{\beta^0_{\min}} \longrightarrow 0,$$

then there exists a solution $\breve{\beta}$ to (4) such that $P(\mathrm{sign}(\breve{\beta}) = \mathrm{sign}(\beta^0)) \longrightarrow 1$, as $\min_{jt} n_{jt} \to \infty$ and $p \to \infty$.

Remark 8. Instead of requiring $\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{(s^2\log p)/\min_{j,t} n_{jt}} = o(1)$, we can use the following weak condition

$$\|(\Sigma_{JJ})^{-1}\|_\infty \cdot \sqrt{\frac{s^2\log p}{\min_{j,t} n_{jt}}} \le \frac{\eta}{Q_1'(4+\eta)},$$

where $Q_1' > Q_1$ is a positive constant. The proof of Theorem 7 is very similar to the proof of Theorem 6. We only show the proof of Theorem 7 briefly in the Appendix.

4. Simulation Study

In this section, we perform numerical studies using simulated examples. We use DISCOM and DISCOM-Huber to denote our proposed methods using sample covariance estimates and Huber's M-estimates, respectively. The proposed methods using the fast tuning parameter selection method shown in Section 2.4 are called Fast-DISCOM and Fast-DISCOM-Huber, respectively. For each example, we compare our proposed methods with (1) Lasso: Lasso method which only uses the samples with complete observations; (2) Imputed-Lasso: Lasso method which uses all samples with missing data imputed by the Soft-thresholded SVD method (Mazumder et al. 2010); (3) Ridge: Ridge regression method which only uses the samples with complete observations; (4) Imputed-Ridge: Ridge regression method which uses all samples with missing data imputed by the Soft-thresholded SVD method; and (5) IMSF (Yuan et al. 2012): the IMSF method which uses all available data without imputing the missing data. We study four simulated examples, where the data are generated from the Gaussian distribution or some heavy-tailed distributions.

For each example, the data are generated from three modalities and each modality has 100 predictors. The training dataset is composed of 100 samples with complete observations, 100 samples with observations from the first and the second modalities, 100 samples with observations from the first and the third modalities, and 100 samples with observations only from the first modality. The tuning dataset contains 200 samples with complete observations and the testing dataset contains 400 samples with complete observations. All methods use the tuning dataset to choose the best tuning parameters. For the four simulated examples, samples with complete observations are generated from the linear model as follows.

Example 1: The predictors $(x_{i1}, x_{i2}, \dots, x_{ip})^T \sim N(0, \Sigma)$ with $\sigma_{jt} = 0.6^{|j-t|}$. The true coefficient vector

$$\beta^0 = (0.5, 0.5, 0.5, \underbrace{0, \cdots, 0}_{97}, 0.5, 0.5, 0.5, \underbrace{0, \cdots, 0}_{97}, 0.5, 0.5, 0.5, \underbrace{0, \cdots, 0}_{97}).$$

The true model is $Y = \mathbf{X}\beta^0 + \boldsymbol{\epsilon}$, where the errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n \overset{\text{iid}}{\sim} N(0, 1)$.

Example 2: The predictors $(x_{i1}, x_{i2}, \dots, x_{ip})^T \sim N(0, \Sigma)$, where Σ is a block diagonal matrix with $p/5$ blocks. Each block is a 5×5 square matrix with ones on the main diagonal and 0.15 elsewhere. The true coefficient vector

$$\beta^0 = (\underbrace{0.5, \cdots, 0.5}_{5}, \underbrace{0, \cdots, 0}_{95}, \underbrace{0.5, \cdots, 0.5}_{5}, \underbrace{0, \cdots, 0}_{95}, \underbrace{0.5, \cdots, 0.5}_{5}, \underbrace{0, \cdots, 0}_{95}).$$

The true model is $Y = \mathbf{X}\beta^0 + \boldsymbol{\epsilon}$, where the errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n \overset{i.i.d}{\sim} N(0, 1)$.

Example 3: The predictors $(x_{i1}, x_{i2}, \dots, x_{ip})^T \sim t_5(\mathbf{0}, 0.6\Sigma)$, where Σ is the same as the covariance matrix shown in Example 1. For this multivariate t-distribution with the degrees of freedom 5, the variances of all predictors are equal to 1. The true coefficient vector β^0 is the same as the vector shown in Example 1. The true model is $Y = \mathbf{X}\beta^0 + \boldsymbol{\epsilon}$, where the errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ follow the Student's t-distribution with degrees of freedom 10.

Example 4: The predictors $(x_{i1}, \dots, x_{ip})^T \sim$ the mixture distribution $\rho \cdot N(\mathbf{0}, 10\mathbf{I}) + (1-\rho) \cdot N(\mathbf{0}, 0.5\mathbf{I})$, where $\rho = 0.03$ and $\mathbf{I}$ is a $p \times p$ identity matrix. The true coefficient vector β^0 is the same as the vector shown in Example 1. The true model is $Y = \mathbf{X}\beta^0 + \boldsymbol{\epsilon}$, where the errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ follow the Skew-t distribution (Azzalini and Capitanio 2013) with degrees of freedom 4.

For each example, we repeated the simulation 30 times. To evaluate different methods, we use the following five measures: ℓ_2 distance $\|\hat{\beta} - \beta^0\|_2$, mean squared error (MSE), false-positive rate (FPR), false-negative rate (FNR), and the elapsed time (in seconds) using R. Tables 1 and 2 show the performance comparison of different methods in the Gaussian case and the heavy-tailed case, respectively. The results indicate that our proposed methods deliver the best performance on all these four examples. For the Gaussian case shown in Table 1, DISCOM delivers better performance than the DISCOM-Huber method. For the heavy-tailed case shown in Table 2, DISCOM-Huber performs better. These numerical results are consistent with our theoretical studies shown in Section 3.

In addition, as shown in Tables 1 and 2, for the Lasso and ridge regression, using the imputed data can improve performance in most cases. However, as shown in Table 1, the Lasso method using the imputed data may deliver worse estimate of the true coefficient vector β^0, possibly due to the block-missing pattern. Compared with the Lasso and Ridge regression methods using the imputed dataset or only the samples with complete observations, the IMSF method delivers better estimation and prediction. On the other hand, IMSF method has high false-positive rates for all four simulated examples. The comparison between IMSF and our proposed DISCOM and DISCOM-Huber shows that our proposed methods could use all available data more effectively and therefore acquires better performance.

For each simulation of the four examples, our proposed Fast-DISCOM method using the fast tuning parameter selection method uses only 4 seconds while our original DISCOM method uses about 13 s. The Fast-DISCOM method is also faster than the IMSF method which uses about 7 seconds for each simulation. On the other hand, we can observe that the computing times of our original DISCOM and DISCOM-Huber methods are still acceptable. For Examples 1 and 2 generated from the Gaussian distribution, although the Fast-DISCOM method does not perform as well as the DISCOM method, it has better estimation, prediction, and model selection performance than the Lasso, ridge regression and IMSF methods. Similarly, for Examples 3 and 4 generated from the heavy-tailed distributions, although the Fast-DISCOM-Huber method does not perform as well as the DISCOM-Huber method, it also has better performance than the Lasso, ridge regression and IMSF methods. These simulation results indicate that our proposed new tuning parameter selection method accelerates the computational speed without sacrificing the estimation, prediction, and model selection performance too much.

5. Real Data Analysis

In this section, we show the analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data as an application example. The main goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography

Table 1. Performance comparison for the Gaussian case.

	Example 1					Example 2				
Methods	$\|\hat{\beta} - \beta^0\|_2$	MSE	FPR	FNR	TIME	$\|\hat{\beta} - \beta^0\|_2$	MSE	FPR	FNR	TIME
Lasso	0.655 (0.026)	1.431 (0.045)	0.069 (0.004)	0.015 (0.009)	0.016 (0.000)	0.920 (0.025)	1.988 (0.059)	0.133 (0.007)	0.002 (0.002)	0.019 (0.004)
Imputed-Lasso	0.674 (0.017)	1.338 (0.018)	0.076 (0.007)	0.004 (0.004)	0.802 (0.006)	0.690 (0.013)	1.546 (0.030)	0.122 (0.007)	0.000 (0.000)	1.099 (0.008)
Ridge	1.270 (0.004)	3.962 (0.062)	1.000 (0.000)	0.000 (0.000)	0.025 (0.000)	1.662 (0.006)	5.262 (0.066)	1.000 (0.000)	0.000 (0.000)	0.025 (0.000)
Imputed-Ridge	1.094 (0.013)	2.304 (0.035)	1.000 (0.000)	0.000 (0.000)	0.780 (0.006)	1.332 (0.009)	3.130 (0.048)	1.000 (0.000)	0.000 (0.000)	1.093 (0.008)
IMSF	0.585 (0.020)	1.358 (0.037)	0.173 (0.009)	0.000 (0.000)	5.554 (0.068)	0.777 (0.016)	1.730 (0.040)	0.291 (0.012)	0.000 (0.000)	5.900 (0.075)
DISCOM	0.416 (0.013)	1.133 (0.016)	0.025 (0.003)	0.000 (0.000)	13.552 (0.078)	0.600 (0.020)	1.378 (0.033)	0.074 (0.007)	0.000 (0.000)	12.391 (0.064)
DISCOM-Huber	0.434 (0.013)	1.145 (0.016)	0.026 (0.003)	0.000 (0.000)	28.618 (0.886)	0.605 (0.021)	1.380 (0.035)	0.076 (0.008)	0.000 (0.000)	25.907 0.122
Fast-DISCOM	0.465 (0.015)	1.160 (0.016)	0.039 (0.005)	0.000 (0.000)	3.600 (0.027)	0.641 (0.017)	1.438 (0.033)	0.109 (0.006)	0.000 (0.000)	3.241 (0.029)
Fast-DISCOM-Huber	0.481 (0.015)	1.173 (0.016)	0.036 (0.004)	0.000 (0.000)	16.802 (0.081)	0.655 (0.020)	1.457 (0.037)	0.100 (0.007)	0.000 (0.000)	16.767 (0.096)

Note: The values in the parentheses are the standard errors of the measures.

Table 2. Performance comparison for the heavy-tailed case.

	Example 3					Example 4				
Methods	$\|\hat{\beta} - \beta^0\|_2$	MSE	FPR	FNR	TIME	$\|\hat{\beta} - \beta^0\|_2$	MSE	FPR	FNR	TIME
Lasso	0.751 (0.036)	1.809 (0.055)	0.070 (0.006)	0.056 (0.017)	0.021 (0.004)	1.305 (0.029)	3.331 (0.087)	0.064 (0.007)	0.419 (0.054)	0.020 (0.005)
Imputed-Lasso	0.751 (0.023)	1.669 (0.039)	0.071 (0.008)	0.026 (0.010)	0.687 (0.010)	0.930 (0.030)	2.699 (0.073)	0.147 (0.014)	0.033 (0.016)	0.530 (0.013)
Ridge	1.294 (0.004)	4.454 (0.114)	1.000 (0.000)	0.000 (0.000)	0.028 (0.003)	1.420 (0.006)	3.548 (0.069)	1.000 (0.000)	0.000 (0.000)	0.039 (0.006)
Imputed-Ridge	1.143 (0.013)	2.731 (0.064)	1.000 (0.000)	0.000 (0.000)	0.657 (0.010)	1.326 (0.011)	3.342 (0.080)	1.000 (0.000)	0.000 (0.000)	0.527 (0.011)
IMSF	0.622 (0.025)	1.637 (0.041)	0.173 (0.013)	0.004 (0.004)	6.569 (0.297)	1.048 (0.028)	2.878 (0.083)	0.189 (0.012)	0.052 (0.017)	6.989 (0.188)
DISCOM	0.579 (0.022)	1.560 (0.038)	0.037 (0.004)	0.004 (0.004)	12.086 (0.134)	0.871 (0.025)	2.590 (0.067)	0.193 (0.017)	0.011 (0.006)	12.362 (0.153)
DISCOM-Huber	0.507 (0.017)	1.452 (0.025)	0.027 (0.003)	0.000 (0.000)	26.073 (0.104)	0.780 (0.021)	2.468 (0.054)	0.137 (0.012)	0.004 (0.004)	26.925 (0.228)
Fast-DISCOM	0.601 (0.022)	1.604 (0.047)	0.040 (0.004)	0.004 (0.004)	3.317 (0.041)	1.151 (0.025)	3.028 (0.079)	0.207 (0.019)	0.085 (0.033)	3.626 (0.050)
Fast-DISCOM-Huber	0.561 (0.021)	1.496 (0.031)	0.035 (0.004)	0.000 (0.000)	17.835 (0.079)	0.786 (0.022)	2.482 (0.055)	0.137 (0.013)	0.000 (0.000)	17.042 (0.134)

[Note: The values in the parentheses are the standard errors of the measures.]

(PET), and some other biological markers and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In our study, we extracted features from three modalities: structural MRI, fluorodeoxyglucose PET, and CerebroSpinal Fluid (CSF). Imaging preprocessing was performed for MRI and PET images. For the MRI, after some correction, spatial segmentation, and registration steps, we obtained the subject lableled image based on the Jacob template (Kabani et al. 1998) with 93 manually labeled regions of interest (ROI). For each of the 93 ROIs in the labeled MRI, we computed the volume of gray matter as a feature. For each PET image, we first aligned the PET

Table 3. Performance comparison for the ADNI data.

	MSE		Number of Features		TIME	
Methods	Mean	SE	Mean	SE	Mean	SE
Lasso	5.711	0.341	11.733	1.638	0.009	0.002
Imputed-Lasso	4.711	0.082	86.700	8.559	0.559	0.017
Ridge	5.273	0.204	191.000	0.000	0.010	0.000
Imputed-Ridge	4.478	0.055	191.000	0.000	0.177	0.006
IMSF	4.630	0.079	28.400	3.025	2.960	0.073
DISCOM	4.285	0.068	27.933	2.261	4.675	0.028
DISCOM-Huber	4.161	0.059	23.100	0.846	10.348	0.025
Fast-DISCOM	4.146	0.055	28.100	0.809	1.565	0.007
Fast-DISCOM-Huber	4.123	0.069	25.833	1.311	8.012	0.019

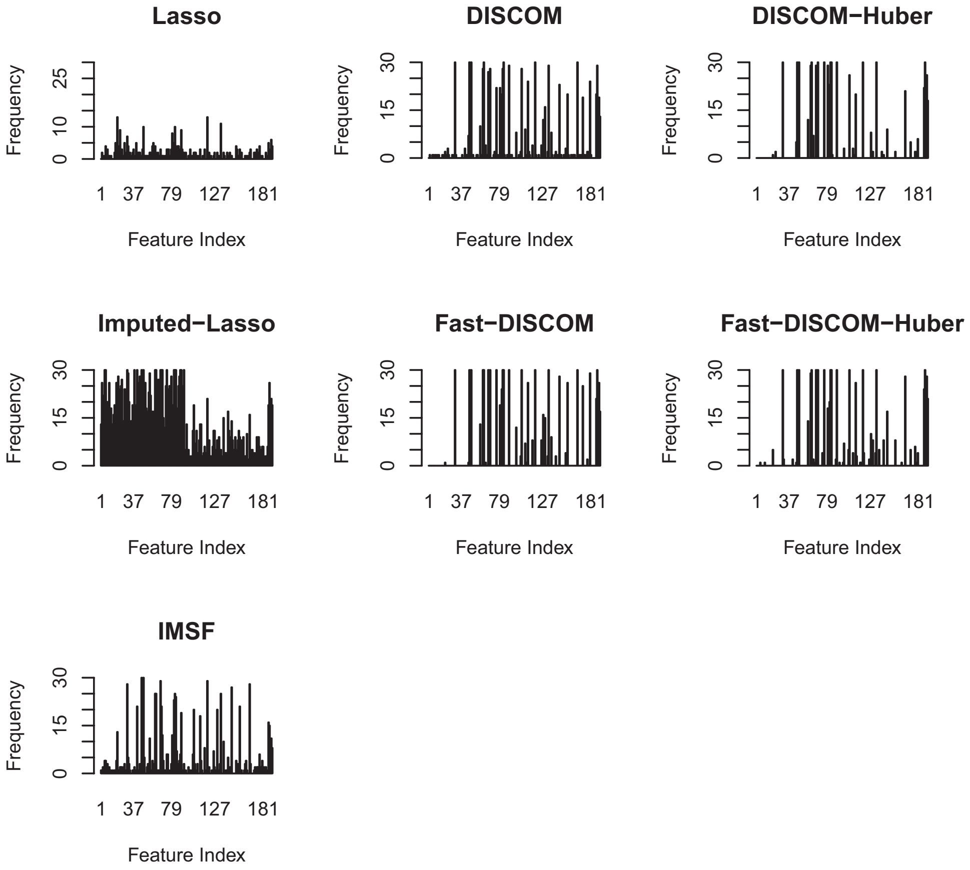

Figure 2. Selection frequency of 191 features for the prediction of MMSE score.

image to its respective MRI using affine registration. Then, we calculated the average intensity of every ROI in the PET image as a feature. Therefore, for each ROI, we have one MRI feature and one PET feature. For the CSF modality, five biomarkers were used in this study, namely amyloid β ($A\beta42$), CSF total tau (t-tau), tau hyperphosphorylated at threonine 181 (p-tau), and two tau ratios with respective to $A\beta42$ (i.e., t-tau/$A\beta42$ and p-tau/$A\beta42$).

After data processing, we got 93 features from MRI, 93 features from PET, and 5 features from CSF. There are 805 subjects in total, including (1) 199 subjects with complete MRI, PET, and CSF features, (2) 197 subjects with only MRI and PET features, (3) 201 subjects with only MRI and CSF features, and (4) 208 subjects with only MRI features. The response variable used in our study is the Mini Mental State Examination (MMSE) score. As a brief 30-point questionnaire test, MMSE can be used to examine a patient's arithmetic, memory and orientation (Folstein et al. 1975). It is very useful to help evaluate the stage of AD pathology and predict future progression. We will use all available data from MRI, PET, and CSF to predict the MMSE score.

In our analysis, we divided the data into three parts: training dataset, tuning dataset, and testing dataset. The training dataset consists of all subjects with incomplete observations and 40 randomly selected subjects with complete MRI, PET, and CSF features. The tuning dataset consists of another 40 randomly selected subjects (different from the training dataset) with complete observations. The testing dataset contains the other 119 subjects with complete observations. The tuning dataset was used to choose the best tuning parameters for all methods and the testing dataset was 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.

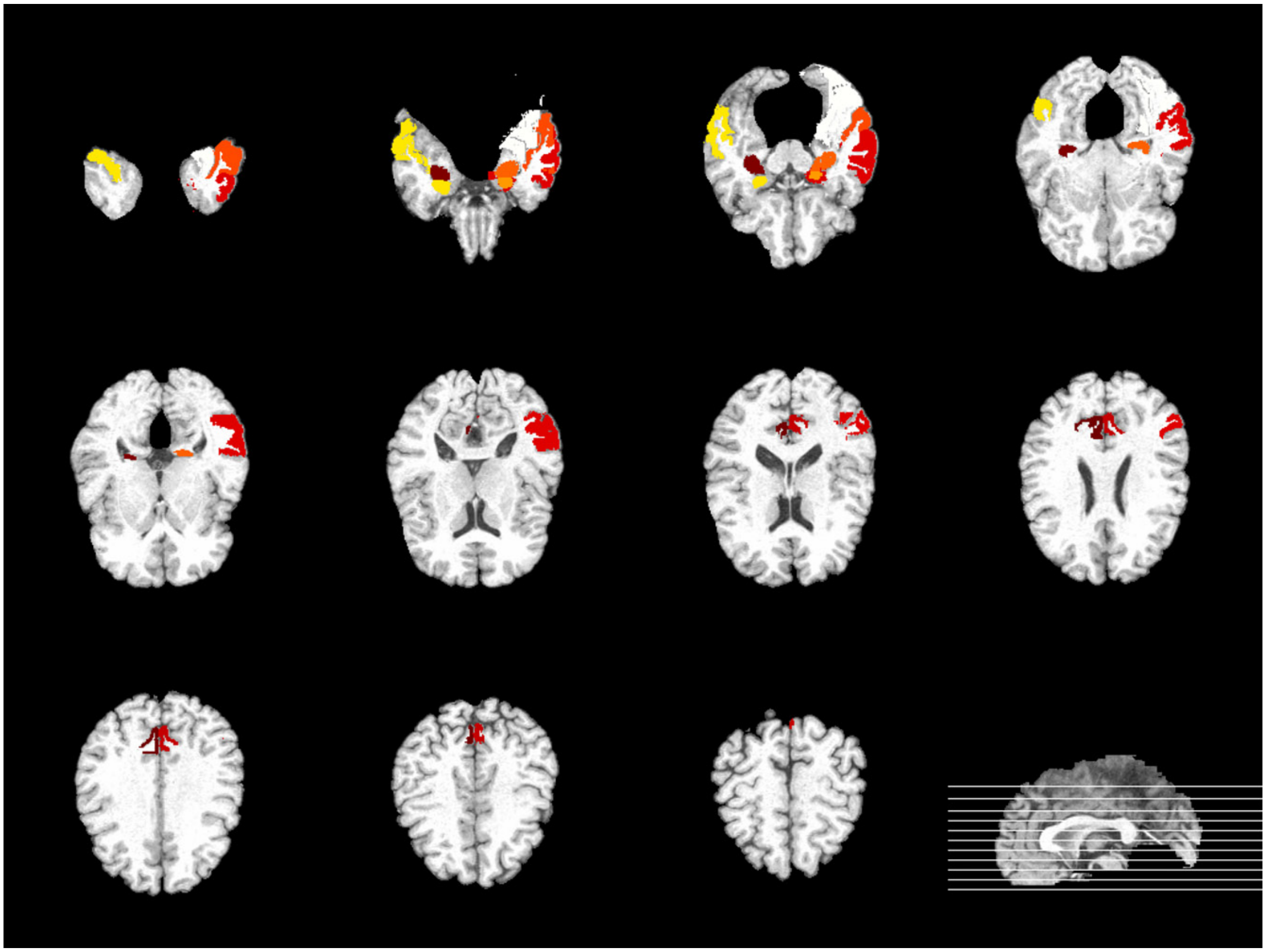

Figure 3. The multi-slice view of the brain regions always selected by DISCOM-Huber and Fast-DISCOM-Huber.

The results in Table 3 show that our proposed Fast-DISCOM-Huber method acquires the best prediction performance. All our proposed DISCOM methods deliver better performance than the Lasso, Ridge, and IMSF methods. The IMSF method has better prediction performance than the Lasso and ridge regression using only samples with complete observations. However, IMSF does not perform as well as the ridge regression using the imputed data. Regarding the model selection, since the number of variables selected by the Lasso is at most the sample size (Zou and Hastie 2005), as shown in Table 3, the Lasso method using the imputed data selected many more features than the method using only samples with complete observations. Both IMSF and our proposed methods could deliver a model with relatively small numbers of features.

Figure 2 shows the selection frequency of all the 191 features. The selection frequency of each feature is defined as the times of being selected in the 30 times replications. As shown in Figure 2, for our proposed DISCOM methods, some features were always selected and many features were never selected in the 30 times replications. This means that our method could deliver relatively robust performance on model selection. However, for some other methods such as the Imputed-Lasso method, they selected very different features in different replications and therefore many features have positive and low selection frequencies. For the Imputed-Lasso method, one possible reason for the unstable performance on model selection is due to the randomness involved in the imputation of a lot of block-missing data.

To further understand our results, since each MRI feature and each PET feature are corresponding to one ROI, we can examine whether the selected features are meaningful by studying their corresponding brain regions. In our 30 times of experiments using different random splits, there are 9 MRI features and 2 PET features always selected by our proposed DISCOM-Huber and Fast-DISCOM-Huber methods. Figure 3 shows the multi-slice view of the brain regions (regions with color) corresponding to these 11 features. Among these 11 brain regions, some regions such as hippocampal formation right (30th region), uncus left (46th region), middle temporal gyrus left (48th region), hippocampus formation left (69th region) and amygdale right (83th region), are known to be highly correlated with AD and MCI by many studies using group comparison methods (Misra et al. 2009; Zhang and Shen 2012). It would be interesting to study whether the other six always selected brain regions are truly related with AD by some scientific experiments.

In addition, as shown in Table 3, all our proposed DISCOM methods solve this real data analysis problem with 191 features within 11 seconds. This indicates that the time cost of our methods is not very expensive. In summary, our real data analysis indicates that our proposed method can solve practical problems well.

6. Conclusion

In this paper, we propose a new two-step procedure to find the optimal linear prediction of a continuous response variable using the block-missing multimodality predictors. In the first step, we estimate the covariance matrix of the predictors using a linear combination of the identity matrix, and 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. We also use all available information to estimate the cross covariance vector between the predictors and the response variable. Robust estimate based on Huber's M-estimate is also proposed for the heavy-tailed case. In the second step, based on the estimated covariance matrix and the cross-covariance vector, a modified Lasso estimator is used to deliver a sparse estimate of the coefficients in the optimal linear prediction. The effectiveness of the proposed method is demonstrated by both theoretical and numerical studies. The comparison between our proposed method and several existing ones also indicates that our method has promising performance on estimation, prediction, and model selection for the block-missing multimodality data.

Acknowledgments

The authors would like to thank the editor, the associate editor, and reviewers, whose helpful comments and suggestions led to a much improved presentation. This research was partially supported by NSF grants IIS1632951, DMS1821231 and NIH grants R01GM126550, P01CA142538.

References

Azzalini, A., and Capitanio, A. (2013), *The Skew-Normal and Related Families*, Cambridge, UK: Cambridge University Press. [9]

Bickel, P. J., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 2577–2604. [2]

Bühlmann, P., and van der Geer, S. (2011), *Statistics for High-Dimensional Data (Springer Series in Statistics)*, New York: Springer. [7]

Bunea, F. (2008), "Consistent Selection via the Lasso for High Dimensional Approximating Regression Models," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, Institute of Mathematical Statistics, Vol. 3, pp. 122–137. [7]

Cai, J., Candès, E. J., and Shen, Z. (2010), "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization*, 20, 1956–1982. [1]

Cai, T., Cai, T. T., and Zhang, A. (2016), "Structured Matrix Completion With Applications to Genomic Data Integration," *Journal of the American Statistical Association*, 111, 621–633. [1]

Cai, T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684. [2]

Cai, T. T., and Zhang, A. (2016), "Minimax Rate-optimal Estimation of High-Dimensional Covariance Matrices With Incomplete Data," *Journal of Multivariate Analysis*, 150, 55–74. [2]

Catoni, O. (2012), "Challenging the Empirical Mean and Empirical Variance: A Deviation Study," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, Vol. 48, pp. 1148–1185. [7]

Datta, A., Zou, H., et al. (2017), "Cocolasso for High-dimensional Error-in-variables Regression," *The Annals of Statistics*, 45, 2400–2426. [2,6,8]

Fan, J., Li, Q., and Wang, Y. (2017), "Estimation of High Dimensional Mean Regression in the Absence of Symmetry and Light Tail Assumptions," *Journal of the Royal Statistical Society*, Series B, 79, 247–265. [8]

Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1361. [1]

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975), "Mini-mental State: A Practical Method for Grading the Cognitive State of Patients for the Clinician," *Journal of Psychiatric Research*, 12, 189–198. [11]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [5]

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. (1999), "Imputing Missing Data for Gene Expression Arrays," Tech. rep., Stanford University. [1]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [4]

Jeng, X. J., and Daye, Z. J. (2011), "Sparse Covariance Thresholding for High-dimensional Variable Selection,' *Statistica Sinica*, 625–657. [2,5,8]

Kabani, N., MacDonald, D., Holmes, C., and Evans, A. (1998), "A 3D Atlas of the Human Brain," *NeuroImage*, 7, S717. [10]

Ledoit, O., and Wolf, M. (2004), "A Well-conditioned Estimator for Large-dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411. [4]

Loh, P.-L., and Wainwright, M. J. (2012), "High-dimensional Regression With Noisy and Missing Data: Provable Guarantees With Nonconvexity," *The Annals of Statistics*, 40, 1637–1664. [2]

Lounici, K. et al. (2014), "High-dimensional Covariance Matrix Estimation With Missing Observations," *Bernoulli*, 20, 1029–1058. [2]

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *The Journal of Machine Learning Research*, 11, 2287–2322. [9]

Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 1436–1462. [7]

Meinshausen, N., and Yu, B. (2009), "Lasso-type Recovery of Sparse Representations for High-dimensional Data," *The Annals of Statistics*, 246–270. [7]

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: Results From ADNI," *NeuroImage*, 44, 1415–1422. [12]

Raskutti, G., Wainwright, M. J., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research*, 11, 2241–2259. [6,7]

——— (2011), "Minimax rates of Estimation for High-dimensional Linear Regression Over ℓ_q-balls," *IEEE Transactions on Information Theory*, 57, 6976–6994. [6]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-dimensional Covariance Estimation by Minimizing ℓ_1-penalized log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [8]

Rothman, A. J. (2012), "Positive Definite Estimators of Large Covariance Matrices," *Biometrika*, 99, 733–740. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Van De Geer, S. A., and Bühlmann, P. (2009), "On the Conditions Used to Prove Oracle Results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392. [6,7]

Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [8]

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., and Ye, J. (2012), "Multi-source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data," *NeuroImage*, 61, 622–632. [2,9]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [7]

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. [12]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [7,8]

Zhou, S., van de Geer, S., and Bühlmann, P. (2009), "Adaptive Lasso for High Dimensional Regression and Gaussian Graphical Modeling," arXiv preprint arXiv:0903.2515. [7]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [7]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1,12]