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

Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu

1Department of Biostatistics, State University of New York at Buffalo; 2Department of Biostatistics, University of North Carolina at Chapel Hill; 3Department of Radiology and BRIC, University of North Carolina at Chapel Hill; 4Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea; 5Department 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.

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 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
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.

![Figure 1](Image) An illustration of block-missing multimodality data with three modalities.
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 $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_\infty$, 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 $\max_i \sum_j |a_{ij}|$, respectively. For a vector $\mathbf{b} \in R^{m \times 1}$, we use $\|\mathbf{b}\|_2$, $\|\mathbf{b}\|_{\infty}$, and $\|\mathbf{b}\|_1$ to denote the $\ell_2$ norm $\sqrt{\sum_i b_i^2}$, the max norm $\max_i |b_i|$, and the $\ell_1$ norm $\sum_i |b_i|$, 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 $k$-th modality. In addition, let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}) = (x_{11}, x_{12}, \ldots, x_{np})^T$ 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 $\Sigma$. We use $\text{cov}(x_i, y_1) = (c_{11}, c_{12}, \ldots, c_{p})^T \in R^p$ to denote the cross-covariance vector between $x_i$ and $y_1$.

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

$$\beta^0 = (\beta_{11}^0, \beta_{12}^0, \ldots, \beta_{p}^0)^T = \arg \min_\beta E \left[ \left( y - \sum_{j=1}^p X_j \beta_j \right)^2 \right] = \Sigma^{-1} C. \quad (1)$$

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

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

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

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

where $\lambda$ 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 $\Sigma$ and the cross-covariance vector $C$. The estimates of $\Sigma$ and $C$ are denoted as $\hat{\Sigma}$ and $\hat{C}$, respectively. This step is very important to make full use of the block-missing multi-modality data. In the second step, we estimate $\beta^0$ by solving the following optimization problem:

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

2.2. Standard Estimates of $\Sigma$ and $C$

Considering block-missing multi-modality 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 $j$th predictor are centered to have mean 0. A natural initial unbiased estimate of $\Sigma$ using all available data is the sample covariance matrix

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

For the block-missing multi-modality data, the above initial estimate $\hat{\Sigma}$ may have negative eigenvalues due to the unequal sample sizes $n_{jt}$’s. Therefore, it is not a good estimate of the covariance matrix $\Sigma$ 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 $\hat{\Sigma}$.

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

$$\hat{\Sigma}_I = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{21} & \cdots & \hat{\Sigma}_{K1} \\ \hat{\Sigma}_{12} & \hat{\Sigma}_{22} & \cdots & \hat{\Sigma}_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{1K} & \hat{\Sigma}_{2K} & \cdots & \hat{\Sigma}_{KK} \end{pmatrix},$$

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

where $\hat{\Sigma}_I$ is called the intra-modality sample covariance matrix which is a $p \times p$ block-diagonal matrix containing $K$ main diagonal blocks of $\Sigma$, and $\hat{\Sigma}_C = \hat{\Sigma} - \hat{\Sigma}_I$ is called the cross-modality sample covariance matrix containing all the off-diagonal blocks of $\hat{\Sigma}$. We also let $\Sigma_I$ and $\Sigma_C$ 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 $\Sigma_l$ than the cross-modality covariance matrix $\Sigma_C$. Intuitively, it is relatively easier to estimate $\Sigma_l$ than $\Sigma_C$. In view of this characteristic of the block-missing multi-modality data and the possible negative eigenvalues of $\hat{\Sigma}$, we propose to use the following estimator:

$$\hat{\Sigma} = \alpha_1 \hat{\Sigma}_l + \alpha_2 \hat{\Sigma}_C + \alpha_3 I_p,$$

where $\alpha_1$, $\alpha_2$ and $\alpha_3$ are three nonrandom weights, and $I_p$ is a $p \times p$ identity matrix. Considering all possible linear combinations, we can find the optimal linear combination $\hat{\Sigma}^* = \alpha_1^* \hat{\Sigma}_l + \alpha_2^* \hat{\Sigma}_C + \alpha_3^* I_p$ whose expected quadratic loss $E[\| \hat{\Sigma}^* - \Sigma \|_F^2]$ is the minimum. The optimal weights $\alpha_1^*$, $\alpha_2^*$ and $\alpha_3^*$ 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:

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

where the weights $\alpha_1$, $\alpha_2$ and $\alpha_3$ are nonrandom. Denote $\gamma^* = \text{tr}(\Sigma)/p$, $\delta_l^2 = E[\| \hat{\Sigma}_l - \Sigma_l \|_F^2]$, $\delta_C^2 = E[\| \hat{\Sigma}_C - \Sigma_C \|_F^2]$, and $\theta^2 = \gamma^* E[I_p - \Sigma_l]$. The optimal weights are

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

In addition, we have

$$E[\| \hat{\Sigma}^* - \Sigma \|_F^2] = \frac{\delta_l^2 \theta^2}{\theta^2 + \delta_l^2} + \frac{\delta_C^2}{\delta_l^2 + \| \Sigma_C \|_F^2} \leq \delta_l^2 + \delta_C^2.$$
2.4. Estimate of $\beta^0$ in the Optimal Linear Prediction

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

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

where $\alpha_1 \in [0, 1], \alpha_2 \in [0, 1]$ are two weights and $\text{tr}(\hat{\Sigma})/p$ is used to estimate $\gamma^*$. In practice, both $\alpha_1 \in [0, 1], \alpha_2 \in [0, 1]$, and $\lambda$ can be chosen by cross-validation or an additional tuning dataset. To guarantee that the estimated covariance matrix $\hat{\Sigma} = \alpha_1 \hat{\Sigma}_I + \alpha_2 \hat{\Sigma}_C + (1 - \alpha_1) \frac{\text{tr}(\hat{\Sigma})}{p} I_p$ is positive semidefinite, we need to choose reasonable $\alpha_1$ and $\alpha_2$ from the set $\{(\alpha_1, \alpha_2) : \alpha_1 \in [0, 1], \alpha_2 \in [0, 1], \lambda \min(\hat{\Sigma}) \geq 0\}$, 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 $\alpha_1$ and $\alpha_2$ should satisfy the conditions $1 - \alpha_1 = O(\sqrt{\log p} / \min_i n_i)$ and $1 - \alpha_2 = O(\sqrt{\log p} / \min_i n_i)$, respectively. Denote $m_1 = \sqrt{\log p} / \min_i n_i$ and $m_2 = \sqrt{\log p} / \min_i n_i$. 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 $\alpha_1$ and $\alpha_2$ 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 $\hat{\Sigma}$, denoted by $\lambda \min(\hat{\Sigma})$, is nonnegative, we can show that $\hat{\Sigma}$ is positive semidefinite for any nonnegative $k_0$. If $\lambda \min(\hat{\Sigma}) < 0$, since the smallest eigenvalue of $\hat{\Sigma}$ satisfies

$$
\lambda \min(\hat{\Sigma}) \geq \lambda \min(\hat{\Sigma} + k_0 \left[ \begin{array}{c}
\lambda \min((m_2 - m_1) \hat{\Sigma}_I + m_1 \frac{\text{tr}(\hat{\Sigma})}{p} I_p)) - m_2 \lambda \min(\hat{\Sigma}) \end{array} \right]),
$$

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

$$
k_0 \geq -\lambda \min(\hat{\Sigma}) / \left[ -m_2 \cdot \lambda \min(\hat{\Sigma}) + \lambda \min((m_2 - m_1) \hat{\Sigma}_I + m_1 \frac{\text{tr}(\hat{\Sigma})}{p} I_p)) \right].
$$

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

$$
k_{\min} = -\lambda \min(\hat{\Sigma}) / \left[ -m_2 \cdot \lambda \min(\hat{\Sigma}) + \lambda \min((m_2 - m_1) \hat{\Sigma}_I + m_1 \frac{\text{tr}(\hat{\Sigma})}{p} I_p)) \right].
$$

For the block-missing multimodality data, since $m_2 \geq m_1 > 0$, we know that the matrix $(m_2 - m_1) \hat{\Sigma}_I + m_1 \frac{\text{tr}(\hat{\Sigma})}{p} I_p$ 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 $\lambda$ rather than searching three parameters $\alpha_1, \alpha_2$ and $\lambda$. In addition, instead of using the eigen decomposition for each parameter combination to check whether $\hat{\Sigma}$ is positive semidefinite, this method only requires two eigendecompositions of the matrices $\hat{\Sigma}$ and $(m_2 - m_1) \hat{\Sigma}_I + m_1 \frac{\text{tr}(\hat{\Sigma})}{p} I_p$ 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 $\lambda$ 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 $X_1, X_2, \ldots, X_p$ to denote the sample means of those $p$ predictors. We also center the observed responses and use $\tilde{Y}$ to denote the sample mean of the response variable. Let $\beta$ denote the estimated regression coefficient vector calculated from the centered data. Our final predictive model is $\hat{Y} = \sum_{j=1}^p X_j^* \tilde{Y}_j$, 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 $\tilde{\sigma}_j = 1$. For Huber’s M-estimator $\hat{\Sigma}$, we redefine $\tilde{\sigma}_j$ to be 1 for each $j$. Let $\beta$ and $\hat{\beta}$ denote the solutions to (4) using the sample covariance and Huber’s M-estimator, respectively. We assume that $\beta^0$ 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 $\beta_{j}^{0\max} = \max_{i \in J} |\beta_{i}^0|$ and $\beta_{j}^{0\min} = \min_{i \in J} |\beta_{i}^0|$. 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:
Theorem 1. Under condition (A1), if \( \min_{j_t} n_{jt} \geq 6 \log p \), there exists two positive constants
\[
\nu_1 = 8\sqrt{6}(1 + 4L^2) \quad \text{and} \quad \nu_2 = 4 \quad \text{such that}
\]
\[
\max_j P \left( |\delta_{jt} - \sigma_{jt}| \geq \nu_1 \sqrt{\frac{\log p}{\min_{jt} n_{jt}}} \right) \leq \frac{\nu_2}{p^2},
\]
\[
P \left( \|\hat{\Sigma} - \Sigma\|_{\infty} \geq \nu_1 \sqrt{\frac{\log p}{\min_{jt} n_{jt}}} \right) \leq \frac{\nu_2}{p}. \]

There exists another two positive constants \( \nu_3 = 16(1 + \frac{L^2}{\min_{\{\text{var}(y), 1\}}} \) max\{\text{var}(y), 1\} \) and \( \nu_4 = 4 \) such that
\[
\max_j P \left( |\hat{\epsilon}_j - c_j| \geq \nu_3 \sqrt{\frac{\log p}{n_{jt}}} \right) \leq \frac{\nu_4}{p^2},
\]
\[
P \left( \|\hat{\Sigma} - C\|_{\infty} \geq \nu_3 \sqrt{\frac{\log p}{\min_{jt} n_{jt}}} \right) \leq \frac{\nu_4}{p}.
\]

Remark 1. In our theoretical study, we assume that the dimension \( p \) goes to infinity as the sample size \( \min_{jt} n_{jt} \) increases. If we further assume that \( \log p / \min_{jt} n_{jt} = o(1) \), the condition \( \min_{jt} n_{jt} > 6 \log p \) is satisfied if the sample size \( \min_{jt} n_{jt} \) is sufficiently large. Then, Theorem 1 shows that \( \|\hat{\Sigma} - \Sigma\|_{\infty} = O_p(\sqrt{\log p / \min_{jt} n_{jt}}) \). The performance of \( \hat{\Sigma} \) depends on the worst case when there are only \( \min_{jt} n_{jt} \) samples to estimate some entries in \( \Sigma \). In addition, the convergence rate of \( \|\hat{\Sigma} - \Sigma\|_{\infty} = O_p(\sqrt{\log p / \min_{jt} n_{jt}}) \). The performance of \( \hat{\Sigma} \) also depends on the worst case when there are only \( \min_{jt} n_{jt} \) 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 \( \|\hat{\Sigma} - \Sigma\|_{\infty} = O_p(\sqrt{\log p / n_{\text{complete}}}) \) and \( \|\hat{\Sigma} - \Sigma\|_{\infty} = O_p(\sqrt{\log p / n_{\text{complete}}}) \), where \( n_{\text{complete}} \) is the number of samples with complete observations. For the block-missing multimodality data, since \( n_{\text{complete}} \) can be much smaller than \( \min_{jt} n_{jt} \) and \( n_{jt} \), 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 \( \|\hat{\beta} - \beta^0\|_2 \).

Theorem 2. Under conditions (A1) and (A2), let \( 1 - \alpha_1 = O\left(\sqrt{\log p / \min_{jt} n_{jt}}\right) \) and \( 1 - \alpha_2 = O\left(\sqrt{\log p / \min_{jt} n_{jt}}\right) \).

If \( s \sqrt{\log p / \min_{jt} n_{jt}} = o(1) \) and we choose \( \lambda = 2\|\hat{\Sigma} - \hat{\Sigma}_0\|_{\infty} \), then we have
\[
\|\hat{\beta} - \beta^0\|_2 = O_p(\sqrt{s\lambda}) = O_p(\sqrt{\alpha_2^2\|\Sigma\|_{\infty}^{1/2} s\sqrt{\log p / \min_{jt} n_{jt}}}).
\]

Remark 2. As shown in the above Theorem 2, we have \( \|\hat{\beta} - \beta^0\|_2 = O_p(\sqrt{s\lambda}) = O_p(\sqrt{\alpha_2^2\|\Sigma\|_{\infty}^{1/2} s\sqrt{\log p / \min_{jt} n_{jt}}}) \).

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 \( Y = X\beta^0 + \epsilon \), where \( \epsilon \sim N(0, \sigma^2 \mathbf{I}) \), then we will use \( \hat{\Sigma} = X^T X / n \) and \( \hat{\Sigma} = X^T Y / n \) to estimate \( \Sigma \) and \( \hat{\Sigma} \), respectively. Therefore, we have \( \|\hat{\Sigma} - \hat{\Sigma}_0\|_{\infty} = \|X^T \epsilon / n\|_{\infty} = O_p(\sqrt{\log p / n}) \), and \( \|\hat{\beta} - \beta^0\|_2 = O_p(\sqrt{\|\Sigma\|_{\infty} \log p / n}) \), which is the minimax \( \ell_2 \)-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 \( \hat{\Sigma} = \hat{\Sigma}_0 \neq X^T X / n \) and \( \|\hat{\Sigma} - \hat{\Sigma}_0\|_{\infty} = O_p(\sqrt{\|\Sigma\|_{\infty} \log p / n}) \). For the block-missing multimodality data, since the Lasso method may only use the data with complete observations, we have \( \|\hat{\beta}_{\text{Lasso}} - \beta^0\|_2 = O_p(\sqrt{\|\Sigma\|_{\infty} \log p / n_{\text{complete}}}) \), whereas, however, in Theorem 2, for our proposed DISCOM estimate \( \beta \), we have \( \|\hat{\beta} - \beta^0\|_2 = O_p(\sqrt{\|\Sigma\|_{\infty} \log p / n_{\text{complete}}}) \).

In practice, the minimum number of samples with available observations from two modalities (\( \min_{jt} n_{jt} \)) can be much larger than the number of samples with complete observations from all modalities (\( n_{\text{complete}} \)). 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 \( s \sqrt{\log p / \min_{jt} n_{jt}} = o(1) \) is used to guarantee that \( \hat{\Sigma} \) satisfies the RE condition with a high probability if the true covariance matrix \( \Sigma \) satisfies the RE condition (A2). Note that many existing sparse linear regression studies focus on the fixed design where the design matrix \( X \) is considered to be fixed and complete. In that case, \( \hat{\Sigma} = X^T X / n \) is assumed to satisfy the RE condition directly. For the general random design, Van De Geer and Bühlmann (2009) showed that \( \hat{\Sigma} = X^T X / n \) satisfies the RE condition as long as the true covariance matrix \( \Sigma \) satisfies the RE condition and \( s \sqrt{\log p / n} = o(1) \). For the special Gaussian random design, by a global analysis of the full random matrix \( \hat{\Sigma} = X^T X / n \) 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 \( n > \text{Constant} \cdot s \sqrt{\log p} \). 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 \( \hat{\Sigma} \neq X^T X / n \) 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-moving 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, 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 \( \max_{1 \leq j \leq p} E(X^t_j) \leq Q^2_1/48 \) and \( E(y^t) \leq Q^2_2 \),

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 fixed designs (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 \( \hat{\Sigma} \) and \( \hat{C} \). Based on the convergence rates of \( \| \hat{\Sigma} - \Sigma \|_{\text{max}} \) and \( \| \hat{C} - C \|_{\text{max}} \), we will show the convergence rate of \( \| \hat{\beta} - \beta^0 \|_2 \).

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

\[
\max_{j,t} P \left( |\hat{\sigma}_{jt} - \sigma_{jt}| \geq Q_1 \sqrt{\log p/n_j t} \right) \leq \frac{2}{p^3},
\]

\[
P \left( \| \hat{\Sigma} - \Sigma \|_{\text{max}} \geq Q_1 \sqrt{\log p/\min j,t n_{jt}} \right) \leq \frac{2}{p}.
\]

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

\[
\max_{j} P \left( |\hat{c}_j - c_j| \geq 8(Q_1 + Q_2) \sqrt{\log p/n_j} \right) \leq \frac{2}{p^3},
\]

\[
P \left( \| \hat{C} - C \|_{\text{max}} \geq 8(Q_1 + Q_2) \sqrt{\log p/\min j,n_{jt}} \right) \leq \frac{2}{p}.
\]

**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 \( \| \hat{\Sigma} - \Sigma \|_{\text{max}} = O_p(\sqrt{(\log p)/\min j,t n_{jt}}) \) and \( \| \hat{C} - C \|_{\text{max}} = O_p(\sqrt{(\log p)/\min j,n_{jt}}) \). 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 \( \hat{\Sigma} \) and \( \hat{C} \) can be wider than the bounds of Huber’s M-estimators \( \Sigma \) and \( C \), respectively.

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

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

\[
P \left( \| \hat{\Sigma} - \Sigma \|_{\text{max}} \geq \frac{d_1}{2T} \sqrt{\frac{p}{\min j,t n_{jt}}} \right) \leq \frac{d_2}{p^{2h - 2}},
\]

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

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

\[
P \left( \| \hat{C} - C \|_{\text{max}} \geq \frac{d_3}{2T} \sqrt{\frac{p}{\min j,n_{jt}}} \right) \leq \frac{d_4}{p^{2h - 2}},
\]

where \( d_3 > 0 \) and \( d_4 > 0 \) are two constants.

**Remark 4.** Under the moment condition, Theorem 4 shows that \( \| \hat{\Sigma} - \Sigma \|_{\text{max}} = O_p(\sqrt{p/\min j,t n_{jt}}) \) and \( \| \hat{C} - C \|_{\text{max}} = O_p(\sqrt{p/\min j,n_{jt}}) \). 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 \( \| \hat{\Sigma} - \Sigma \|_{\text{max}} \) and \( \| \hat{C} - C \|_{\text{max}} \) can be much larger than the bounds of \( \| \hat{\Sigma} - \Sigma \|_{\text{max}} \) and \( \| \hat{C} - C \|_{\text{max}} \), respectively. This necessitates the usage of a robust estimator.

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

**Theorem 5.** Under conditions (A2) and (A3), let \( 1 - \alpha_1 = O(\sqrt{(\log p)/\min j,n_{jt}}), 1 - \alpha_2 = O(\sqrt{(\log p)/\min j,t n_{jt}}) \),

\[
H_{jt} = \frac{Q_1}{12} \sqrt{n_j t \log p} \quad \text{and} \quad H_j = (Q_1 + Q_2)\sqrt{n_j \log p}.
\]

If \( s(\sqrt{(\log p)/\min j,t n_{jt}}) = o(1) \) and let \( \lambda = 2\| \hat{\Sigma} - \Sigma \beta^0 \|_{\text{max}} \), then we have \( \| \hat{\beta} - \beta^0 \|_2 = O_p(\sqrt{\lambda} = O_p(\sqrt{s(\log p)/\min j,t n_{jt}}) \).\]

**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
Remark 6. If we assume that \( p \) is fixed, for the sub-Gaussian case considered in Section 3.1, we can show that \( \| \hat{\Sigma} - \Sigma \|_{\infty} = O_p(\min_{j \in J} n_j^{-1/2}) \) and \( \| \hat{C} - C \|_{\infty} = O_p((\min_{j \in J} n_j)^{-1/2}) \) 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 \( \| \hat{\Sigma} - \Sigma \|_{\infty} = O_p((\min_{j \in J} n_j)^{-1/2}) \) and \( \| \hat{C} - C \|_{\infty} = O_p((\min_{j \in J} n_j)^{-1/2}) \) 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 \( \| \hat{\beta} - \beta^0 \|_2 = O_p(\sqrt{\log n} \frac{1}{\min_{j \in J} n_j}) \) and \( \| \hat{C} - C \|_{\infty} = O_p((\min_{j \in J} n_j)^{-1/2}) \). 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.

\[
(A4) \quad \| \Sigma_{J^c} \Sigma_{J^c}^{-1} \|_{\infty} \leq 1 - \eta, \quad \text{where} \quad \eta \in (0, 1) \quad \text{is a constant},
\]

\( \Sigma_{J^c} \) is the sub-matrix of \( \Sigma \) with row indices in the set \( J^c \) and column indices in the set \( J \) and \( \Sigma_{J^c} \) is the sub-matrix of \( \Sigma \) 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 \( 1 - \alpha_1 = O(\sqrt{\log p} / \min_{j \in J} n_j) \) and \( 1 - \alpha_2 = O(\sqrt{\log p} / \min_{j \in J} n_j) \). If

\[
\| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{\frac{s \log p}{\min_{j \in J} n_j}} \rightarrow 0,
\]

\[
1 + \frac{\sqrt{\log p}}{\min_{j \in J} n_j} \rightarrow 0, \quad \lambda \cdot \| (\Sigma_{J^c})^{-1} \|_{\infty} \rightarrow 0,
\]

then there exists a solution \( \hat{\beta} \) to (4) such that \( P(\text{sign}(\hat{\beta}) = \text{sign}(\beta^0)) \rightarrow 1 \), as \( \min_{j \in J} n_j \rightarrow \infty \) and \( p \rightarrow \infty \).

**Remark 7.** Note that the condition \( \| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{\frac{s \log p}{\min_{j \in J} n_j}} = o(1) \) is used to guarantee that (a) \( \| (\hat{\Sigma}_{J^c})^{-1} \|_{\infty} \leq \text{Constant} \cdot \| (\Sigma_{J^c})^{-1} \|_{\infty} \) and (b) \( \| \hat{\Sigma}_{J^c}^{-1} \|_{\infty} \leq 1 - \eta' \) if \( \| \hat{\Sigma}_{J^c}^{-1} \|_{\infty} \leq 1 - \eta \) 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 \( \hat{\Sigma} = \hat{X}^T \hat{X}/n \) 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 \( \hat{X} \neq X^T X/n \) for the block-missing multi-modality data, we use the condition \( \| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{(s \log p) / \min_{j \in J} n_j} = o(1) \) 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 \( \| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{s \log p} / \min_{j \in J} n_j = o(1) \), we can use the following weak condition

\[
\| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot s \log p / \min_{j \in J} n_j = o(v_1),
\]

where \( v_1 > v_1 \) is a positive constant.

**Theorem 7.** Under conditions (A3) and (A4), let \( H_j = Q_1 \sqrt{\log p} / \min_{j \in J} n_j \), \( H_j = O(1 + Q_2) \sqrt{\log p} / \min_{j \in J} n_j \), and \( \alpha_1 = O((\sqrt{\log p} / \min_{j \in J} n_j), 1 - \alpha_2 = O((\sqrt{\log p} / \min_{j \in J} n_j)) \). If

\[
\| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{(s \log p) / \min_{j \in J} n_j} \rightarrow 0,
\]

then there exists a solution \( \hat{\beta} \) to (4) such that \( P(\text{sign}(\hat{\beta}) = \text{sign}(\beta^0)) \rightarrow 1 \), as \( \min_{j \in J} n_j \rightarrow \infty \) and \( p \rightarrow \infty \).

**Remark 8.** Instead of requiring \( \| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot \sqrt{(s \log p) / \min_{j \in J} n_j} = o(1) \), we can use the following weak condition

\[
\| (\Sigma_{J^c})^{-1} \|_{\infty} \cdot s \log p / \min_{j \in J} n_j \leq Q_1(4 + \eta),
\]

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 third 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 \((x_{i1}, x_{i2}, \ldots, x_{ip})^T \sim N(0, \Sigma)\) with \(\sigma_{ij} = 0.6^{\lvert i-j \rvert}\). The true coefficient vector

\[
\beta^0 = (0.5, 0.5, 0.5, 0, \ldots, 0, 0.5, 0.5, 0, \ldots, 0, 0, 0, 0, \ldots, 0, 0, 0, 0, \ldots, 0).
\]

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

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

\[
\beta^0 = (0.5, \ldots, 0.5, 0, \ldots, 0, 0.5, \ldots, 0.5, 0, \ldots, 0, 0.5, \ldots, 0.5, 0, \ldots, 0, 0.5, \ldots, 0.5, 0, \ldots, 0).
\]

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

**Example 3.** The predictors \((x_{i1}, x_{i2}, \ldots, x_{ip})^T \sim t_5(0, 0.6\Sigma)\), where \(\Sigma\) 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 \(\beta^0\) is the same as the vector shown in Example 1. The true model is \(Y = X\beta^0 + \epsilon\), where the errors \(\epsilon_1, \epsilon_2, \ldots, \epsilon_n \) follow the Student’s \(t\)-distribution with degrees of freedom 10.

**Example 4.** The predictors \((x_{i1}, \ldots, x_{ip})^T \sim \text{the mixture distribution } p \cdot N(\mathbf{0}, 10\mathbf{I}) + (1 - p) \cdot N(\mathbf{0}, 0.5\mathbf{I})\), where \(p = 0.03\) and \(\mathbf{I}\) is a \(p \times p\) identity matrix. The true coefficient vector \(\beta^0\) is the same as the vector shown in Example 1. The true model is \(Y = X\beta^0 + \epsilon\), where the errors \(\epsilon_1, \epsilon_2, \ldots, \epsilon_n \) 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: \(\ell_2\) distance \(\lVert \hat{\beta} - \beta^0 \rVert_2\), 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 \(\beta^0\), 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$|\hat{\beta} - \beta^0|_2$</th>
<th>MSE</th>
<th>FPR</th>
<th>FNR</th>
<th>TIME</th>
<th>$|\hat{\beta} - \beta^0|_2$</th>
<th>MSE</th>
<th>FPR</th>
<th>FNR</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>0.655 (0.026)</td>
<td>1.431</td>
<td>0.069</td>
<td>0.015</td>
<td>0.016</td>
<td>0.920 (0.025)</td>
<td>1.988</td>
<td>0.133</td>
<td>0.002</td>
<td>0.019</td>
</tr>
<tr>
<td>Imputed-Lasso</td>
<td>0.674 (0.017)</td>
<td>1.338</td>
<td>0.076</td>
<td>0.004</td>
<td>0.002</td>
<td>0.690 (0.013)</td>
<td>1.546</td>
<td>0.122</td>
<td>0.000</td>
<td>1.099</td>
</tr>
<tr>
<td>Ridge</td>
<td>1.270 (0.004)</td>
<td>3.962</td>
<td>1.000</td>
<td>0.000</td>
<td>0.025</td>
<td>1.662 (0.006)</td>
<td>5.262</td>
<td>1.000</td>
<td>0.000</td>
<td>0.025</td>
</tr>
<tr>
<td>Imputed-Ridge</td>
<td>1.094 (0.013)</td>
<td>2.304</td>
<td>1.000</td>
<td>0.000</td>
<td>0.780</td>
<td>1.332 (0.009)</td>
<td>3.130</td>
<td>1.000</td>
<td>0.000</td>
<td>1.093</td>
</tr>
<tr>
<td>IMSF</td>
<td>0.585 (0.020)</td>
<td>1.358</td>
<td>0.173</td>
<td>0.000</td>
<td>5.554</td>
<td>0.777 (0.016)</td>
<td>1.730</td>
<td>0.291</td>
<td>0.000</td>
<td>5.900</td>
</tr>
<tr>
<td>DISCOM</td>
<td>0.416 (0.013)</td>
<td>1.133</td>
<td>0.025</td>
<td>0.000</td>
<td>13.552</td>
<td>0.600 (0.020)</td>
<td>1.378</td>
<td>0.074</td>
<td>0.000</td>
<td>12.391</td>
</tr>
<tr>
<td>DISCOM-Huber</td>
<td>0.434 (0.013)</td>
<td>1.145</td>
<td>0.026</td>
<td>0.000</td>
<td>28.618</td>
<td>0.605 (0.021)</td>
<td>1.380</td>
<td>0.076</td>
<td>0.000</td>
<td>25.907</td>
</tr>
<tr>
<td>Fast-DISCOM</td>
<td>0.465 (0.015)</td>
<td>1.160</td>
<td>0.039</td>
<td>0.000</td>
<td>3.600</td>
<td>0.641 (0.017)</td>
<td>1.438</td>
<td>0.109</td>
<td>0.000</td>
<td>3.241</td>
</tr>
<tr>
<td>Fast-DISCOM-Huber</td>
<td>0.481 (0.015)</td>
<td>1.173</td>
<td>0.026</td>
<td>0.000</td>
<td>16.802</td>
<td>0.655 (0.020)</td>
<td>1.457</td>
<td>0.100</td>
<td>0.000</td>
<td>16.767</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>$|\hat{\beta} - \beta^0|_2$</th>
<th>MSE</th>
<th>FPR</th>
<th>FNR</th>
<th>TIME</th>
<th>$|\hat{\beta} - \beta^0|_2$</th>
<th>MSE</th>
<th>FPR</th>
<th>FNR</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>0.751 (0.036)</td>
<td>1.809</td>
<td>0.070</td>
<td>0.015</td>
<td>0.021</td>
<td>1.305 (0.029)</td>
<td>3.331</td>
<td>0.064</td>
<td>0.419</td>
<td>0.020</td>
</tr>
<tr>
<td>Imputed-Lasso</td>
<td>0.751 (0.036)</td>
<td>1.669</td>
<td>0.071</td>
<td>0.026</td>
<td>0.687</td>
<td>0.930 (0.030)</td>
<td>2.699</td>
<td>0.147</td>
<td>0.033</td>
<td>0.530</td>
</tr>
<tr>
<td>Ridge</td>
<td>1.294 (0.004)</td>
<td>4.454</td>
<td>1.000</td>
<td>0.000</td>
<td>0.028</td>
<td>1.420 (0.006)</td>
<td>3.548</td>
<td>1.000</td>
<td>0.000</td>
<td>0.039</td>
</tr>
<tr>
<td>Imputed-Ridge</td>
<td>1.143 (0.013)</td>
<td>2.731</td>
<td>1.000</td>
<td>0.000</td>
<td>0.657</td>
<td>1.326 (0.011)</td>
<td>3.342</td>
<td>1.000</td>
<td>0.000</td>
<td>0.527</td>
</tr>
<tr>
<td>IMSF</td>
<td>0.622 (0.025)</td>
<td>1.637</td>
<td>0.173</td>
<td>0.004</td>
<td>6.569</td>
<td>1.048 (0.028)</td>
<td>2.878</td>
<td>0.189</td>
<td>0.052</td>
<td>6.989</td>
</tr>
<tr>
<td>DISCOM</td>
<td>0.579 (0.022)</td>
<td>1.560</td>
<td>0.037</td>
<td>0.004</td>
<td>12.086</td>
<td>0.871 (0.025)</td>
<td>2.590</td>
<td>0.193</td>
<td>0.011</td>
<td>12.362</td>
</tr>
<tr>
<td>DISCOM-Huber</td>
<td>0.507 (0.017)</td>
<td>1.452</td>
<td>0.027</td>
<td>0.000</td>
<td>26.073</td>
<td>0.780 (0.021)</td>
<td>2.468</td>
<td>0.137</td>
<td>0.004</td>
<td>26.925</td>
</tr>
<tr>
<td>Fast-DISCOM</td>
<td>0.601 (0.022)</td>
<td>1.604</td>
<td>0.040</td>
<td>0.004</td>
<td>3.317</td>
<td>1.151 (0.025)</td>
<td>3.028</td>
<td>0.207</td>
<td>0.085</td>
<td>3.626</td>
</tr>
<tr>
<td>Fast-DISCOM-Huber</td>
<td>0.561 (0.021)</td>
<td>1.496</td>
<td>0.035</td>
<td>0.000</td>
<td>17.835</td>
<td>0.786 (0.022)</td>
<td>2.482</td>
<td>0.137</td>
<td>0.000</td>
<td>17.042</td>
</tr>
</tbody>
</table>

[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 labeled 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.

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

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 $\beta$ ($A\beta$42), CSF total tau (t-tau), tau hyperphosphorylated at threonine 181 (p-tau), and two tau ratios with respective to $A\beta$42 (i.e., t-tau/$A\beta$42 and p-tau/$A\beta$42).

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.
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