Published in final edited form as:
Stat Anal Data Min. 2012 December 1; 5(6): . doi:10.1002/sam.11158.

Multiple Response Regression for Gaussian Mixture Models with Known Labels

Wonyul Lee, **Ying Du**, **Wei Sun**, **D. Neil Hayes**, and **Yufeng Liu***
University of North Carolina at Chapel Hill

Abstract

Multiple response regression is a useful regression technique to model multiple response variables using the same set of predictor variables. Most existing methods for multiple response regression are designed for modeling homogeneous data. In many applications, however, one may have heterogeneous data where the samples are divided into multiple groups. Our motivating example is a cancer dataset where the samples belong to multiple cancer subtypes. In this paper, we consider modeling the data coming from a mixture of several Gaussian distributions with known group labels. A naive approach is to split the data into several groups according to the labels and model each group separately. Although it is simple, this approach ignores potential common structures across different groups. We propose new penalized methods to model all groups jointly in which the common and unique structures can be identified. The proposed methods estimate the regression coefficient matrix, as well as the conditional inverse covariance matrix of response variables. Asymptotic properties of the proposed methods are explored. Through numerical examples, we demonstrate that both estimation and prediction can be improved by modeling all groups jointly using the proposed methods. An application to a glioblastoma cancer dataset reveals some interesting common and unique gene relationships across different cancer subtypes.

Keywords

Covariance estimation; GLASSO; Hierarchical penalty; LASSO; Multiple response; Regression; Sparsity

1 Introduction

Multivariate regression with a univariate response is a common and popular technique that builds a model to predict a response variable given a set of predictor variables. In the statistical learning literature, many multivariate regression techniques are designed for the setting of a univariate response case. In many applications, one may have multiple response variables available. A simple approach is to regress each response variable separately on the same set of explanatory variables. Although it is simple and popular, the univariate response approach ignores the joint information among response variables.

To improve estimation and prediction accuracy, it can be advantageous to model the response variables jointly. Breiman and Friedman [1] proposed a method called the “curd and whey”, which was shown to achieve improved prediction accuracy over univariate regression techniques when there are correlations among the response variables. To achieve variable selection, Turlach et al. [2] proposed a penalized method using the max-L_1 penalty. Their method aims to select a common subset that can be used as predictors for all response

*Address for correspondence: Yufeng Liu, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, CB3260, University of North Carolina, Chapel Hill, NC 27599. yfliu@email.unc.edu.

variables. Yuan et al. [3] proposed another shrinkage method to model response variables jointly. Their idea is to obtain dimension reduction by encouraging sparsity among singular values of the regression coefficient matrix. The embedded dimension reduction is very useful especially when the dimension of predictor variables is much higher than the sample size.

Besides the regression coefficient matrix, it can also be useful to estimate the conditional inverse covariance matrix. Under the Gaussian assumption, this matrix closely relates to Gaussian graphical models and provides useful interpretation of the relationship among response variables. Recently, Rothman et al. [4] and Lee and Liu [5] proposed to model response variables jointly in a penalized likelihood framework using L_1 regularization. In particular, they assume that the conditional distribution of response variables given predictors is multivariate Gaussian. Under this assumption, they perform joint estimation of the regression coefficient matrix and the conditional inverse covariance matrix of response variables given predictors. In the estimation procedure, Lee and Liu [5] used weighted L_1 penalties on both matrices in order to encourage sparsity among the entries of the estimated matrices. Their results indicate that simultaneous modeling of the multiple response variables can provide more accurate estimation of both regression coefficients and the inverse covariance matrix.

The previous work mentioned above assumes that all observations come from a single multivariate Gaussian distribution. However, in some applications, this assumption can be too strong. For example, we consider the glioblastoma multiforme (GBM) cancer data set studied by The Cancer Genome Atlas (TCGA) Research Network [6]. Verhaak et al. [7] showed that the GBM patients can be divided into four subtypes based on their gene expressions. Based on their study, gene expressions of patients within each subtype are very similar. However, patients in different subtypes can be very different from each other. Therefore, the assumption of one multivariate Gaussian distribution for all patients may not be valid. In this paper, we consider modeling the data arisen from a mixture of several Gaussian distributions. Specifically, we model gene expression data of the patients of a particular subtype by a multivariate Gaussian distribution, which can vary from one subtype to another. Here we assume that the Gaussian mixture labels are given. To tackle this problem, a naive approach is to model each group separately. However, this approach ignores the common structures that may exist across different groups. Therefore, it might be more useful to model all groups jointly so that the common structures can be estimated from the aggregated data.

In this paper, we propose three approaches to model all groups jointly via penalizing parameter matrices together. The first two approaches are plug-in methods and the third one is to estimate all parameter matrices jointly. In particular, for the first approach, we plug in a reasonable estimator of the inverse covariance matrices to estimate the regression coefficient matrices. For the second approach, we estimate the inverse covariance matrices instead after plugging in a good estimator of the regression coefficient matrices. The last approach simultaneously estimates the regression coefficient matrices and the inverse covariance matrices. These methods are penalized log-likelihood approaches with the multivariate mixture Gaussian assumption.

In the following sections, we describe the new proposed methods in more details with theoretical justification and numerical examples. In Section 2, we introduce our proposed methods. Section 3 explores their theoretical properties. Section 4 develops computational algorithms to obtain solutions for proposed methods. Simulated examples are presented in Section 5 to demonstrate performance of our methods and Section 6 provides analysis of a

glioblastoma cancer data example. We conclude the paper with some discussion in Section 7. The proofs of the theorems are provided in the Appendix.

2 Methodology

Consider the dataset with G different groups. Suppose the g-th group contains n_g observations of p covariates and m response variables. Let $\mathbf{y}_i^{(g)}=(y_{i1}^{(g)},\ldots,y_{im}^{(g)})^T$; $i = 1, \ldots, n_g$, be m-dimensional responses and $\mathbf{Y}^{(g)}=[\mathbf{y}_1^{(g)},\ldots,\mathbf{y}_{n_g}^{(g)}]^T$ be the $n_g \times m$ response matrix in the g-th group. Let $\mathbf{x}_i^{(g)}=(x_{i1}^{(g)},\ldots,x_{ip}^{(g)})^T$; $i = 1, \ldots, n_g$, be p-dimensional predictors and $\mathbf{X}^{(g)}=[\mathbf{x}_1^{(g)},\ldots,\mathbf{x}_{n_g}^{(g)}]^T$ be the $n_g \times p$ design matrix in the g-th group. Consider the multiple response linear regression model in the g-th group,

$$\mathbf{Y}^{(g)}=\mathbf{X}^{(g)}\mathbf{B}^{(g)}+\mathbf{e}^{(g)}, \quad \text{with} \quad \mathbf{e}^{(g)}=[\varepsilon_1^{(g)},\ldots,\varepsilon_{n_g}^{(g)}]^T,$$

where $\mathbf{B}^{(g)}=\{\beta_{jk}^{(g)}\}$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, is an unknown $p \times m$ parameter matrix. The errors $\varepsilon_i^{(g)}=(\varepsilon_{i1}^{(g)},\ldots,\varepsilon_{im}^{(g)})^T$; $i = 1, \ldots, n_g$, are i.i.d. m-dimensional random vectors following a multivariate normal distribution $\mathbf{N}(0, \mathbf{\Sigma}^{(g)})$ with a nonsingular covariance matrix $\mathbf{\Sigma}^{(g)}$. Let $\mathbf{C}^{(g)}=(\sum^{(g)})^{-1}=(c_{j,j'}^{(g)})_{m\times m}$; $j, j' = 1, \ldots, m$.

Our goal is to estimate $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ so that we can perform prediction and achieve graphical interpretation among response variables, where $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\} = \{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)}), g = 1, \ldots, G\}$. The most direct way to estimate $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ is to build G individual maximum likelihood models. More specifically, the maximum likelihood estimator of $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ can be obtained via maximizing the following conditional log-likelihoods on $\mathbf{X}^{(g)}$,

$$\frac{n_g}{2}\log\det(\mathbf{C}^{(g)})-\frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}, \quad g=1,\ldots,G, \quad (1)$$

up to a constant not depending on $(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})$. It is well-known that the resulting estimator of $\mathbf{B}^{(g)}$ is the ordinary least squares estimator and it does not make use of the joint information among response variables. To incorporate the joint information among response variables in the estimation procedure, we can apply the method proposed by Lee and Liu [5]. In particular, the estimator is given by solving

$$\underset{\mathbf{B}^{(g)},\mathbf{C}^{(g)}}{\operatorname{argmin}}\left\{-l(\mathbf{B}^{(g)},\mathbf{C}^{(g)})+\lambda_1\sum_{j,k}\left|\beta_{jk}^{(g)}\right|+\lambda_2\sum_{s\neq t}\left|c_{st}^{(g)}\right|\right\}, \quad (2)$$

where $l(\mathbf{B}^{(g)}, \mathbf{C}^{(g)}) = n_g \log\det(\mathbf{C}^{(g)})-\mathrm{tr}\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\}$; $g = 1, \ldots, G$.

Motivated from the technique for a single linear model as in (2), we consider penalization for (1) to improve estimation. In particular, estimation of $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ can be improved if some common information across groups can be shared in the estimation procedure. Note that the optimization problem suggested in (2) can be solved individually within each group.

Therefore, it does not utilize the common information across groups. However, since these groups may have shared information with similar structure, it can be useful to consider the connection.

In this section, we propose methods that combine G individual models to improve prediction and estimation. Our goal is to estimate $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ simultaneously to identify the common and unique structures across groups. Note that there are two parameter matrices in each group, $\mathbf{B}^{(g)}$ and $\mathbf{C}^{(g)}$, involved in the estimation, and in many applications only one of them is of main interest. Hence, we consider three different approaches, two plug-in methods and one joint method. In Sections 2.1 and 2.2, we introduce two different plug-in penalized likelihood methods, one is for multiple response regression and the other one is for inverse covariance estimation. In the plug-in method for multiple response regression, we estimate $\{\mathbf{C}^{(g)}\}$ first and then plug it into the likelihood to estimate $\{\mathbf{B}^{(g)}\}$ using penalization. In the plug-in method for inverse covariance estimation, we estimate $\{\mathbf{B}^{(g)}\}$ first and then incorporate the information to estimate $\{\mathbf{C}^{(g)}\}$. In Section 2.3, we estimate $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ together via double penalization.

2.1 Plug-in Hierarchical LASSO estimator

Our goal in this section is to estimate the regression coefficients $\{\mathbf{B}^{(g)}\}$, while assuming the inverse covariance estimates $\{\hat{\mathbf{C}}^{(g)}\}$ is available. Although $\mathbf{B}^{(g)}$ can be different for different g, we expect they share some common structures. In particular, for our cancer application example, different groups correspond to patients with different subtypes of brain cancer. Thus, patients from different groups are likely to have a lot of similarities although there are important differences among various subtypes. This motivates us to perform joint estimation of $\{\mathbf{B}^{(g)}\}$ through shrinkage. It is desirable to identify the common and unique structure on $\{\mathbf{B}^{(g)}\}$ through the penalty.

Suppose we have the inverse covariance estimates, $\{\hat{\mathbf{C}}^{(g)}\}$, available. Let $\beta_{jk}=(\beta_{jk}^{(1)},\ldots,\beta_{jk}^{(G)})^T$. Regression parameters, $(\beta_{jk}^{(1)},\ldots,\beta_{jk}^{(G)})$, corresponding to the same response variable and the same predictor variable are treated as a group. We consider a new penalized likelihood method, namely the plug-in hierarchical LASSO (PHL) estimator, to estimate $\{\mathbf{B}^{(g)}\}$ by solving

$$\underset{(\mathbf{B}^{(g)})_{g=1}^{G}}{\arg\min} \sum_{g=1}^{G} \operatorname{tr}\left\{(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})\hat{\mathbf{C}}^{(g)}(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}+\lambda_1\sum_{j,k}p(\beta_{jk}),$$
$$\text{where}\quad p(\beta_{jk})=\left(\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|\right)^{1/2}. \tag{3}$$

Here λ_1 is a tuning parameter. The penalty in (3) was proposed by Zhou and Zhu [8], which they call the hierarchical group penalty. This penalty controls the sparsity of $\{\mathbf{B}^{(g)}\}$ hierarchically. As the first level of the hierarchical sparsity, the estimator of $\boldsymbol{\beta}_{jk}$ tends to shrink to a zero vector with the hierarchical penalty as a group if all coefficients in the group are small in magnitude. For the second level of the hierarchical sparsity, if $\boldsymbol{\beta}_{jk}$ is estimated as a nonzero vector, within the group, some coefficients can be still shrunk to zero according to their magnitude. Zhou and Zhu [8] showed the penalty in (3) encourages such a hierarchical sparsity. Intuitively, note that $p(\boldsymbol{\beta}_{jk})$ can be approximated by $\sum_{g=1}^{G}\frac{1}{2(\sum_{g=1}^{G}|\beta_{jk}^{(g),*}|)^{1/2}}\left|\beta_{jk}^{(g)}\right|$ where $\beta_{jk}^{(g),*}$ is close to the solution of (3). Therefore, all coefficients in $\boldsymbol{\beta}_{jk}$ have the same

weight, $\frac{1}{2(\sum_{g=1}^{G}|\beta_{jk}^{(g),*}|)^{1/2}}$, as a group while each coefficient has different amount of penalty according to its magnitude. As a remark, we would like to point out that $p(\beta_{jk})$ for each group serves as a group penalty which encourages group shrinkage. Similar idea was previously considered by Turlach et al. [2], Yuan and Lin [9], Zhang et al. [10], and Zhao et al. [11].

For the procedure in (3), we need to first estimate $\{\mathbf{C}^{(g)}\}$. To that end, we obtain initial estimates of $\{\mathbf{B}^{(g)}\}$ by applying univariate regression techniques within each group. Let $\{\hat{\mathbf{B}}^{(g),0}\}$ be initial estimates. Define $\mathbf{S}^{(g)}$ by

$$\mathbf{S}^{(g)}=\frac{1}{n_g}(\boldsymbol{Y}^{(g)}-\mathbf{X}\hat{\mathbf{B}}^{(g),0})\,(\boldsymbol{Y}^{(g)}-\mathbf{X}\hat{\mathbf{B}}^{(g),0})^{T}. \quad (4)$$

Then $\{\hat{\mathbf{C}}^{(g)}\}$ can be obtained by solving

$$\underset{\mathbf{C}^{(g)}}{\operatorname{argmin}}\left\{-\operatorname{logdet}(\mathbf{C}^{(g)})+\operatorname{tr}(\mathbf{S}^{(g)}\mathbf{C}^{(g)})+\lambda_2\sum_{j\neq k}v_{jk}\,|c_{jk}|\right\},\quad g=1,\ldots,G. \quad (5)$$

This problem is essentially the same as the problem of estimating the inverse covariance matrix in the context of sparse Gaussian graphical models. This technique was considered by many authors [12, 13, 14, 15]. Among various existing algorithms, we adapt the Graphical LASSO (GLASSO) algorithm proposed by Friedman et al. [14].

2.2 Plug-in Hierarchical Graphical LASSO estimator

In Section 2.1, we considered a plug-in method, PHL, which estimates $\{\mathbf{C}^{(g)}\}$ first and then estimates $\{\mathbf{B}^{(g)}\}$ given $\{\hat{\mathbf{C}}^{(g)}\}$. In this section, we propose another plug-in method using $\{\hat{\mathbf{B}}^{(g)}\}$ to estimate $\{\mathbf{C}^{(g)}\}$. In particular, we first estimate $\{\mathbf{B}^{(g)}\}$ by using univariate regression techniques and then obtain $\{\mathbf{S}^{(g)}\}$, defined as

$\mathbf{S}^{(g)}=\frac{1}{n_g}(\boldsymbol{Y}^{(g)}-\mathbf{X}\hat{\mathbf{B}}^{(g)})\,(\boldsymbol{Y}^{(g)}-\mathbf{X}\hat{\mathbf{B}}^{(g)})^{T}$. With $\{\mathbf{S}^{(g)}\}$ available, we propose a penalized likelihood method, the plug-in hierarchical graphical LASSO (PHGL) estimator, by solving

$$\underset{(\mathbf{C}^{(g)})_{g=1}^{G}g=1}{\operatorname{argmin}}\sum_{g=1}^{G}\left\{-n_g\operatorname{logdet}(\mathbf{C}^{(g)})+n_g\operatorname{tr}(\mathbf{S}^{(g)}\mathbf{C}^{(g)})\right\}+\lambda_2\sum_{s\neq t}\left(\sum_{g=1}^{G}\left|c_{st}^{(g)}\right|\right)^{1/2}, \quad (6)$$

where λ_2 is a tuning parameter.

This approach is closely related to the method previously considered by Guo et al. [16]. They considered the problem of estimating the inverse covariance matrix of $\mathbf{Y}^{(g)}$. However, we estimate the conditional inverse covariance matrix of $\mathbf{Y}^{(g)}$ given $\mathbf{X}^{(g)}$. Even though the optimization problem in (6) is technically the same as that in their method, our resulting estimator has different graphical interpretations.

2.3 Doubly Penalized Sparse Estimator

In Sections 2.1 and 2.2, we considered two plug-in methods for estimation of $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$. In this section, we propose to estimate $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ simultaneously. We would like to incorporate the information among different response variables in estimation of $\{\mathbf{B}^{(g)}\}$ and

encourage all groups to share some common structure among $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$. We propose a joint penalized method, the doubly penalized sparse estimator (DPS), by solving

$$\underset{(\mathbf{B}^{(g)},\mathbf{C}^{(g)})_{g-1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G}\left\{-l_g(\mathbf{B}^{(g)},\mathbf{C}^{(g)})\right\}+\lambda_1\sum_{jk}\left(\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|\right)^{1/2}+\lambda_2\sum_{s\neq t}\left(\sum_{g=1}^{G}\left|c_{st}^{(g)}\right|\right)^{1/2}, \quad (7)$$

where $l_g(\mathbf{B}^{(g)}, \mathbf{C}^{(g)}) = n_g \log\det(\mathbf{C}^{(g)}) - \mathrm{tr}\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\}$. As a group penalty, the first penalty term in (7) encourages the hierarchical sparsity among $\{\mathbf{B}^{(g)}\}$. In the meantime, the second penalty term in (7) serves as a group penalty for $\{\mathbf{C}^{(g)}\}$.

Note that the objective function in (7) is not convex with respect to $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$ and the optimization can be unstable when $\max\{n_1, \ldots, n_G\} < p$. With diagonal $\{\mathbf{C}^{(g)}\}$, the first term in $l_g(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})$ can dominate the other terms in the objective function if the trace terms are zero, which may occur when $\max\{n_1, \ldots, n_G\} < p$. Therefore, the objective function can keep decreasing as the diagonal entries in $\{\mathbf{C}^{(g)}\}$ continue to increase. As a result, the numerical solution of $\{\mathbf{C}^{(g)}\}$ in (7) can have large diagonal entries. Such kinds of solutions are not desirable in practice since it implies that the residual variances of response variables are very small. Therefore, if $\max\{n_1, \ldots, n_G\} < p$, the plug-in methods in Sections 2.1 and 2.2 are recommended and can often perform better than the DPS method.

2.4 Model Selection

In Sections 2.1 – 2.3, we proposed two plug-in methods and one joint method for estimation of $\{(\mathbf{B}^{(g)}, \mathbf{C}^{(g)})\}$. To apply these methods, we first need to select the tuning parameters λ_1 and λ_2 in (3), (6), and (7), which control the sparsity of the resulting estimators. For the tuning parameter selection, K-fold cross-validation can be combined to our methods. In particular, the K-fold cross-validation method randomly splits the dataset into K parts of equal sizes. Denote the data in the k-th segment by $\{(\mathbf{X}_{(k)}^{(g)}, \mathbf{Y}_{(k)}^{(g)})\}$. For any given λ_1, λ_2 and k, we estimate the regression coefficient matrices and the inverse covariance matrices using all data except the data in the k-th part and denote them by $\{(\widehat{\mathbf{B}}_{\lambda_1,(-k)}^{(g)}, \widehat{\mathbf{C}}_{\lambda_2,(-k)}^{(g)})\}$. For the PHL method, the optimal tuning parameter $\hat{\lambda_1}$ is selected which minimizes the prediction error defined by

$$\mathrm{CV}(\lambda_1)=\sum_{k=1}^{K}\sum_{g=1}^{G}\left|\left|\boldsymbol{Y}_{(k)}^{(g)}-\mathbf{X}_{(k)}^{(g)}\widehat{\mathbf{B}}_{\lambda_1,(-k)}^{(g)}\right|\right|_F^2, \quad (8)$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. For the PHGL method, we select the optimal tuning parameter $\hat{\lambda_2}$ which maximizes the predictive log-likelihood defined by

$$\mathrm{CV}(\lambda_2)=\sum_{k=1}^{K}\sum_{g=1}^{G}\left[n_{(g,k)}\log\det(\widehat{\mathbf{C}}_{\lambda_2,(-k)}^{(g)})-\mathrm{tr}\left\{(\boldsymbol{Y}_{(k)}^{(g)}-\mathbf{X}_{(k)}^{(g)}\widehat{\mathbf{B}}^{(g)})\widehat{\mathbf{C}}_{\lambda_2,(-k)}^{(g)}(\boldsymbol{Y}_{(k)}^{(g)}-\mathbf{X}_{(k)}^{(g)}\mathbf{B}^{(g)})^T\right\}\right], \quad (9)$$

where $n_{(g,k)}$ is the sample size of the g-th group in the k-th segment. In the DPS method, we first choose the optimal $\hat{\lambda_1}$ using (8) with a prespecified λ_2 and the optimal $\hat{\lambda_2}$ is selected using (9) with the selected $\hat{\lambda_1}$. It helps to avoid a two dimensional grid search of (λ_1, λ_2). We have found in simulations that within a certain range for λ_2, the choice of particular value for λ_2 has very little effect on the optimal $\hat{\lambda_1}$. In particular, in all simulated examples in Section 5, the selected optimal values of λ_1 are almost identical for any prespecified value of

λ_2 within interval $[2^{-6}, 2^6]$. To prespecify a value of λ_2, the optimal tuning parameter in the PHGL method can be used.

The tuning parameters can be also selected using validation sets. In particular, we split the dataset into the training set and the validation set. With given λ_1 and λ_2, we construct the corresponding models by applying our methods to the training set. By using the validation set as $\{(\mathbf{X}_{(k)}^{(g)}, \boldsymbol{Y}_{(k)}^{(g)})\}$ in (8) and (9), we can compute the prediction error and the predictive log-likelihood on this set to select tuning parameters. The cross-validation method is computationally more intensive than using validation sets. We used validation sets for our simulated examples and the 5-fold cross-validation for the glioblastoma cancer data example.

3 Asymptotic Properties

In this section, we investigate the asymptotic behavior of our three proposed methods when sample sizes go to infinity. In particular, we show that the resulting estimators of all three methods satisfy consistency and sparsity with proper choices of tuning parameters. To this end, we use the set-up of Fan and Li [17], Yuan and Lin [12] and Zou [18]. The technical derivation uses the results in Knight and Fu [19]. Without loss of generality, we assume that $n = n_1 = \cdots = n_G$ and n goes to infinity. Define a vector operator for any matrix $A = [a_1, \ldots, a_p]$ by $\mathrm{Vec}(A)=(a_1^T, \ldots, a_p^T)^T$. Let $\boldsymbol{\beta}^*$ = (Vec($\mathbf{B}^{*,(1)}$)^T, …, Vec($\mathbf{B}^{*,(G)}$)^T)^T be the true regression parameter vector and $\boldsymbol{c}^*$ = (Vec($\mathbf{C}^{*,(1)}$)^T, …, Vec($\mathbf{C}^{*,(G)}$)^T)^T be the vector of the entries in the true inverse covariance matrices. The following theorem shows the $\sqrt{n}$-consistency and the sparsity of the solution in (3).

Theorem 1—Suppose that $\lambda_1 n^{-\frac{1}{2}} \to 0$ as $n \to \infty$ and $\hat{\mathbf{C}}^{(g)}$ in (3) is a consistent estimator of $\mathbf{C}^{*,(g)}$; g = 1, …, G. Furthermore, suppose that $\frac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)} \to A^{(g)}$ as $n \to \infty$ where $A^{(g)}$ is a positive definite matrix; g = 1, …, G.

1. (Consistency) There exists a local minimizer of (3) such that $\left|\left|\widehat{\beta}-\beta^*\right|\right| = O_p(\frac{1}{\sqrt{n}})$, where $\boldsymbol{\beta}\widehat{} = (Vec(\mathbf{B}\widehat{^{(1)}})^T, \ldots, Vec(\mathbf{B}\widehat{^{(G)}})^T)^T$;
2. (Sparsity) If $\lambda_1 n^{-\frac{1}{4}} \to \infty, \lim_n P(\widehat{\beta}_{jk}^{(g)}=0)=1$ if $\beta_{jk}^{*,(g)}=0$.

Theorem 1 states that with a consistent estimator of $\mathbf{C}^{*,(g)}$, the PHL estimator is $\sqrt{n}$-consistent. Furthermore, it can identify the true subset of predictor variables asymptotically with probability tending to 1. Similar asymptotic properties hold for the PHGL estimator as stated in the following theorem.

Theorem 2—Suppose that $\lambda_2 n^{-\frac{1}{2}} \to 0$ as $n \to \infty$ and $\mathbf{B}^{\widehat{(g)}}$ in (6) is a $\sqrt{n}$-consistent estimator of $\mathbf{B}^{*,(g)}$; g = 1, …, G.

1. (Consistency) There exists a local minimizer of (6) such that $\left|\left|\hat{\mathbf{c}}-\boldsymbol{c}^*\right|\right| = O_p(\frac{1}{\sqrt{n}})$, where $\hat{\mathbf{c}}$ = ($Vec(\hat{\mathbf{C}}^{(1)})^T$, …, $Vec(\hat{\mathbf{C}}^{(G)})^T$)^T;
2. (Sparsity) If $\lambda_2 n^{-\frac{1}{4}} \to \infty, \lim_n P(\hat{c}_{jk}^{(g)}=0)=1$ if $c_{jk}^{*,(g)}=0$.

In theorems 1 and 2, we establish the consistency and sparsity of plug-in estimators. The following theorem shows the similar asymptotic properties of the DPS solution in which $\{\mathbf{B}^{\widehat{(g)}}\}$ and $\{\hat{\mathbf{C}}^{(g)}\}$ are obtained together.

Theorem 3—Suppose that $\lambda_1 n^{-\frac{1}{2}} \to 0$ and $\lambda_2 n^{-\frac{1}{2}} \to 0$ as $n \to \infty$. In addition to that, suppose that $\frac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)} \to A^{(g)}$ as $n \to \infty$ where $A^{(g)}$ is a positive definite matrix; $g = 1, \ldots, G$.

1. (Consistency) There exists a local minimizer of (7) such that

$$\left|\left|(\widehat{\beta}^T, \widehat{\mathbf{c}}^T)^T - (\beta^{*T}, \boldsymbol{c}^{*T})^T\right|\right| = O_p\left(\frac{1}{\sqrt{n}}\right),$$

   where $\widehat{\boldsymbol{\beta}} = (Vec(\mathbf{B}^{\widehat{(1)}})^T, \ldots, Vec(\mathbf{B}^{\widehat{(G)}})^T)^T$ and $\widehat{\mathbf{c}} = (Vec(\widehat{\mathbf{C}}^{(1)})^T, \ldots, Vec(\widehat{\mathbf{C}}^{(G)})^T)^T$;

2. (Sparsity of $\{\mathbf{B}^{\widehat{(g)}}\}$) If $\lambda_1 n^{-\frac{1}{4}} \to \infty, \lim_n P(\widehat{\beta}_{jk}^{(g)} = 0) = 1$ if $\beta_{jk}^{*,(g)} = 0$;

3. (Sparsity of $\{\widehat{\mathbf{C}}^{(g)}\}$) If $\lambda_2 n^{-\frac{1}{4}} \to \infty, \lim_n P(\widehat{c}_{jk}^{(g)} = 0) = 1$ if $c_{jk}^{*,(g)} = 0$.

4 Computational Algorithm

In this section, we describe computational algorithms to solve problems (3), (6), and (7). In particular, we apply the coordinate-descent algorithms used by Lee and Liu [5] iteratively, with combination of the local linear approximation (LLA) [20].

We now describe the algorithm for the PHL method in details. Denote the estimates of $\beta_{jk}^{(g)}$ from the i-th iteration by $(\widehat{\beta}_{jk}^{(g)})^{(i)}$. Then by applying the LLA, the penalty term in (3) at the (i + 1)-th iteration can be approximated as follows,

$$p(\beta_{jk}) = \left(\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|\right)^{1/2} \approx \frac{\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|}{2\left(\sum_{g=1}^{G}\left|(\widehat{\beta}_{jk}^{(g)})^{(i)}\right|\right)^{1/2}}.$$

Then, at the (i + 1)-th iteration, the problem (3) is decomposed into G individual optimization problems

$$\underset{\mathbf{B}^{(g)}}{\operatorname{argmin}} \operatorname{tr}\left\{(\boldsymbol{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\widehat{\mathbf{C}}^{(g)}(\boldsymbol{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\} + \lambda_1 \sum_{j,k} w_{jk}\left|\beta_{jk}^{(g)}\right|, \quad (10)$$

where $w_{jk} = \frac{1}{2}\left(\sum_{g=1}^{G}\left|(\widehat{\beta}_{jk}^{(g)})^{(i)}\right|\right)^{-1/2}$ and $g = 1, \ldots, G$. The optimization problem (10) is exactly the problem of estimating the regression parameter matrix with the plug-in inverse covariance matrix. It can be solved by applying the coordinate-descent algorithm for the plug-in weighted LASSO method proposed by Lee and Liu [5]. The basic idea of the coordinate-descent algorithm is to optimize each parameter at one time with other parameters being fixed at the current solution. Therefore, the algorithm for (3) proceeds as follows:

Algorithm for the PHL Method

Step 1 (Initial value). Set the separate LASSO solution $\{(\mathbf{B}^{(g)})^{(i)}\}$; $g = 1, \ldots, G$ as the initial value for $\{\mathbf{B}^{(g)}\}$.

Step 2 (Updating rule). For $g = 1, \ldots, G$, update $\{(\mathbf{B}^{(g)})^{(i)}\}$ by applying the coordinate-descent algorithm for the the plug-in weighted LASSO method [5] to the problem (10).

Step 3 (Iteration). Repeat Step 2 until convergence.

Next we describe the algorithm for the PHGL method in Section 2.2. Similar to the algorithm for the PHL method, we first apply the LLA to the objective function in (6) with the current estimates $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$. Then, at the $(i + 1)$-th iteration, the problem (6) is decomposed into G individual optimization problems

$$\underset{\mathbf{C}^{(g)}}{\operatorname{argmin}} \left\{ -n_g \operatorname{logdet}(\mathbf{C}^{(g)}) + n_g \operatorname{tr}(\mathbf{S}^{(g)}\mathbf{C}^{(g)}) \right\} + \lambda_2 \sum_{s \neq t} v_{st} \left| c_{st}^{(g)} \right|, \quad (11)$$

where $v_{st} = \frac{1}{2}\left(\sum_{g=1}^{G} \left| (\hat{c}_{jk}^{(g)})^{(i)} \right| \right)^{-1/2}$ and $g = 1, \ldots, G$. The problem (11) can be solved by applying the GLASSO algorithm. Therefore, the algorithm for (6) proceeds as follows:

Algorithm for the PHGL Method

Step 1 (Initial value). Set the separate GLASSO solution $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$; $g = 1, \ldots, G$ as the initial value for $\{\mathbf{C}^{(g)}\}$.

Step 2 (Updating rule). For $g = 1, \ldots, G$, update $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$ by applying the GLASSO algorithm to the problem (11).

Step 3 (Iteration). Repeat Step 2 until convergence.

Next, we combine the above two algorithms to solve problem (7) for the doubly penalized method, DPS. The algorithm can be summarized as follows:

Algorithm for the DPS Method

Step 1 (Initial values of $\{\mathbf{B}^{(g)}\}$ and $\{\mathbf{C}^{(g)}\}$). Set the separate LASSO solution $\{(\mathbf{B}^{(g)})^{(i)}\}$ as the initial value for $\{\mathbf{B}^{(g)}\}$ and the separate GLASSO solution $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$ as the initial value of $\{\mathbf{C}^{(g)}\}$.

Step 2 ($\{\mathbf{B}^{(g)}\}$ updating rule). For a given $\{(\mathbf{B}^{(g)})^{(i)}\}$, update $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$ by applying the algorithm for the PHGL method.

Step 3 ($\{\mathbf{C}^{(g)}\}$ updating rule). For a given updated $\{(\hat{\mathbf{C}}^{(g)})^{(i)}\}$, update $\{(\mathbf{B}^{(g)})^{(i)}\}$ by applying the algorithm for the PHL method.

Step 4 (Iteration). Repeat Steps 2 and 3 until convergence.

As we point out in Section 2.3, when $\max\{n_1, \ldots, n_G\} < p$, the solution can possibly be unstable with very small residual variances. In that case, the plug-in methods may perform better.

5 Simulated Examples

In this section, simulation studies are carried out to assess the performance of our proposed methods. In particular, we compare our proposed methods with several existing methods. All five methods are described below.

Method 1 (M1). We model each group separately. In particular, we apply the doubly penalized maximum likelihood (DML) method by Lee and Liu [5] separately to each group. The estimator is given by solving (2). This method will be referred as DML1.

Method 2 (M2). In this approach, all groups are combined into one dataset as if they come from a common Gaussian distribution. We apply the DML method to the combined dataset. We name this method as DML2.

Method 3 (M3). We first estimate $\{\mathbf{B}^{(g)}\}$ by applying LASSO to each response variable separately in each group. Once we have an estimator of $\{\mathbf{B}^{(g)}\}$, we compute residuals and apply GLASSO to estimate $\{\mathbf{C}^{(g)}\}$. In particular, the estimator of $\{\mathbf{C}^{(g)}\}$ is given by solving (5). The resulting estimator of $\{\mathbf{B}^{(g)}\}$ will be called the LASSO estimator and the resulting estimator of $\{\mathbf{C}^{(g)}\}$ will be referred as the GLASSO estimator.

Method 4 (M4). An initial estimate of $\{\mathbf{B}^{(g)}\}$ is obtained by applying LASSO. With the initial estimate of $\{\mathbf{B}^{(g)}\}$, we apply our proposed plug-in method, PHGL, to estimate $\{\mathbf{C}^{(g)}\}$ jointly. Once we have the estimator of $\{\mathbf{C}^{(g)}\}$, another plug-in method, PHL, is applied to obtain the final estimate of $\{\mathbf{B}^{(g)}\}$.

Method 5 (M5). We model all groups jointly by applying our proposed method, DPS. In this approach, we estimate both $\{\mathbf{B}^{(g)}\}$ and $\{\mathbf{C}^{(g)}\}$ simultaneously.

Note that Methods 1 and 3 model all groups separately and Method 2 does not allow any unique structure to each group. On the other hand, our proposed methods (Methods 4 and 5) model all groups jointly while allowing unique structures to each group.

We set $G = 3$, $p = 20$ and $m = 20$. For each group, we generate training sets, validation sets, and testing sets with the the same size of $n = 40$. Each data set is generated as follows. First, we produce $\mathbf{B}$ and $\mathbf{C}$ common in all groups. Figure 1 shows the common structure across groups. We create unique structures to each group by adding additional nonzero parameters to each group. In particular, for each $\mathbf{B}^{(g)}$, we randomly pick zero entries and replace them with values randomly chosen from the interval [1,3]. For each $\mathbf{C}^{(g)}$, we randomly pick zero entries and make them have values randomly chosen from interval $[-1, -0.5] \cup [0.5, 1]$. We define ρ as the ratio of the number of unique nonzero entries to the number of common nonzero entries. We consider two values of ρ. The case of $\rho = 0$ does not allow unique structure to each group. The second case has $\rho = 0.25$. Finally, $\mathbf{y}_i^{(g)}$ is generated from $\mathbf{N}(\mathbf{B}^{(g)T}\mathbf{x}_i^{(g)}, \mathbf{C}^{(g)})$, where $\mathbf{x}_i^{(g)}$; $i = 1, \ldots, n$ are i.i.d vectors from $\mathbf{N}(0, I_p)$.

To assess prediction performance, we use the prediction error defined as,

$$\mathrm{PE} = \frac{1}{nmG}\sum_{g=1}^{G}\left|\left|\boldsymbol{Y}^{(g)} - \widehat{\boldsymbol{Y}}^{(g)}\right|\right|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

To compare performance in the estimation of $\{\mathbf{C}^{(g)}\}$, we report the average entropy loss and the average Frobenius loss which are defined as,

$$\mathrm{EL}=\frac{1}{G}\sum_{g=1}^{G}\left[\mathrm{tr}(\sum^{(g)}\widehat{\mathbf{C}}^{(g)})-\log(\left|\sum^{(g)}\widehat{\mathbf{C}}^{(g)}\right|)-m\right],$$

$$\mathrm{FL}=\frac{1}{G}\sum_{g=1}^{G}\left|\left|\mathbf{C}^{(g)}-\widehat{\mathbf{C}}^{(g)}\right|\right|_F^2/\left|\left|\mathbf{C}^{(g)}\right|\right|_F^2.$$

Table 1 and Figures 2–4 summarize the results. When $\rho = 0$, M2 outperforms the others in both prediction and estimation of $\{\mathbf{C}^{(g)}\}$. This is expected because M2 assumes all groups come from the same distribution and that assumption is valid when $\rho = 0$. Therefore, by combining all groups, M2 has more information than other methods. Note that our proposed methods, M4 and M5, also show competitive performance in prediction. When $\rho = 0.25$, M5, one of our proposed methods, shows the best performance in all criteria. This implies that modeling all groups jointly can help us improve both prediction and the estimation of $\{\mathbf{C}^{(g)}\}$ when all groups share some common structures.

Table 2 summarizes the relative computational times of M4 and M5 compared with that of M3. In terms of computational complexity, M5 is more intensive than the other methods while M4 shows competitive computational time. For instance, when $\rho = 0.25$, the computational time of M5 is 30.09 times of that for M3. M5 is computationally more intensive as it estimates all parameter matrices simultaneously. However, in terms of performance, M5 outperforms M3 in both prediction and estimation of $\{\mathbf{C}^{(g)}\}$ in our simulated examples.

6 Application on the Glioblastoma Cancer Data

In this section, we apply our proposed methods to a glioblastoma multiforme (GBM) cancer dataset. In our application, there are 17814 genes and 534 micro-RNAs of 482 GBM patients. The patients were classified into 4 gene expression-based subtypes, namely, Classical, Mesenchymal, Neural, and Proneural with sample sizes of 127, 145, 85 and 125 respectively [7]. One important goal is to regress genes on micro-RNAs to investigate the effect of micro-RNAs on gene expressions. The other goal is to estimate the conditional inverse covariance matrix of gene expressions given micro-RNAs. This matrix can help us to interpret the conditional relationship among genes given micro-RNAs.

To proceed with the analysis, preprocessing is necessary. There are many possibilities for preprocessing. For example, Bair and Tibshirani [21] developed some procedures that utilize both gene expression data and clinical data to select a list of genes for identifying cancer subtypes. In our analysis, the preprocessing step proceeds as follows. Verhaak et al. [7] established 840 signature genes which are highly distinctive for four subtypes. We use these 840 signature genes to explore distinctive effects of micro-RNAs on them. Our proposed methods are needed for genes having correlated residuals. Therefore, to apply our proposed methods, the genes are first grouped into several gene modules with genes more related to each other within each module. Then our proposed methods are applied to each module separately. This approach is sensible for our methodology as a gene module is a set of genes which are closely related. To detect such gene modules, we perform the weighted gene co-expression network analysis (WGCNA) by Zhang and Horvath [22]. WGCNA detects modules using a hierarchical clustering method with the topological overlap dissimilarity measure [23]. Zhang and Horvath [22] pointed out that WGCNA can detect biologically meaningful modules.

By performing WGCNA with the 840 signature genes, we found 14 modules with 60 genes per module on average. It turns out that one of them is particularly interesting as many genes

in the module such as *EGFR* and *PDGFA* are involved in cell proliferation. Moreover, Verhaak et al. [7] demonstrated the essential roles of these genes in GBM tumor genesis. Therefore, we focus on that module hereafter. In particular, there are 90 genes in this module. Among them, we choose top 40 genes with largest Median Absolute Deviations (MADs) since for the 50 genes with low MADs, all regression coefficients are estimated nearly zeros which do not provide any meaningful interpretation. We also select a subset of micro-RNAs which are predicted to target at least one of the selected genes and have large MADs. As a result, 40 genes and 50 micro-RNAs are used for the results in this analysis.

We consider four approaches to estimate the regression coefficient matrices and the residual inverse covariance matrices. In the first approach, we assume that the Gaussian distributions in all subtypes are the same. Therefore, all subtypes are combined into one data set and we apply the doubly penalized maximum likelihood (DML) method by Lee and Liu [5] to the combined data. In particular, the estimator can be obtained by solving (2). In the second approach, we apply LASSO and GLASSO. The detailed description of this approach is presented in M3 in Section 5. In the third approach, we apply our proposed plug-in methods, PHL and PHGL. The last approach uses the DPS method in which all matrices are jointly estimated. The third and fourth approaches can help us to discover the common and unique structures to each group.

For performance assessment, we randomly divide the data set of each subgroup into a training set of size 70 and a test set of the remainder. The tuning parameters are selected using 5-fold cross-validation as discussed in Section 2.4. We perform the random splitting 100 times. By using the test set, we assess prediction performance of several methods including our proposed methods.

Table 3 shows average PE of 100 replications. Note that the DPS, PHL and LASSO methods outperform the DML method. It implies that a single Gaussian assumption for all subtypes might not be reasonable. The LASSO gives comparable, but slightly better prediction accuracy than our PHL and DPS methods. One potential reason is that we allow different tuning parameter values for each response in the LASSO. The more flexible tuning may help the LASSO give slightly better PE.

Figure 5 shows the averaged estimated regression coefficients over 100 replications of several micro-RNAs for some selected genes. In order to produce the heatmap, the DPS estimates are used. The results show some interesting relationships between genes and micro-RNAs that are specific to certain GBM subtypes. For instance, we have observed a negative correlation between *miR222* and its predicted target *GLI2* in the Mesenchymal subtype. *GLI2* is an essential transcription factor mediating cytokine expression in cancer cells [24]. It has been shown that the knockdown of *GLI2* mRNA has significantly decreased the migratory ability of human glioblastoma cells [25]. Herein, our results suggest that the accelerated inflammatory response observed in the GBM Mesenchymal subtype might be partially through miR222-dependent GLI2 regulation [7].

Another example is the anti-correlation between *miR130b* and its predicted target *ARAP2 (CENTD1)* in the GBM Neural subtype. This subtype is typically associated with the gene ontology (GO) categories such as neuron projection and axon and synaptic transmission. Yoon et al. [26] have reported that *ARAP2* associates with focal adhesions and functions downstream of *RhoA* to regulate focal adhesion dynamics in glioblastoma cells. Consistent with this report, our findings suggest that *miR130b* regulates *ARAP2* specifically in the neural subtype.

Additionally, we have observed the subtype-specific correlation between micro-RNAs and non-target genes, indicating an indirect regulation between the two. For instance, our results have identified distinct *EGFR-miR21* correlations in different subtypes. Several research papers have shown that *EGFR* regulates *miR21* in a couple of cancers, including human glioblastoma and lung cancers [27, 28]. Here our observation further indicates that this regulation is subtype-specific in GBM. In the Neural subtype, there was positive correlation between *EGFR* and *miR21* while negative correlations are observed in the subtypes, Messenchymal and Proneural.

Figure 6 shows the estimated conditional inverse covariance structure of genes given micro-RNAs. This structure is obtained from the model using our proposed DPS method. Black edges represent the common structure shared among all subgroups while grey edges represent unique structures to some subgroups. Verhaak et al. [7] claimed that *FGFR3*, *PDGFA* and *EGFR* are all Classical genes in sense that they tend to be highly expressed only in Classical subtype. Thus, it is expected that they have some connectivity among them. However, in Figure 6 from our results, none of them are connected for all subtypes. This implies that in all subtypes, they can be conditionally independent given other genes once we take out the effects of given 50 micro-RNAs on them even though they are marginally correlated. Therefore, joint modeling of all subtypes using our DPS method can help us to interpret similarities and differences of the conditional gene relationships given micro-RNAs among different cancer subtypes.

7 Discussion

In this paper, we propose three methods for modeling several groups jointly to estimate both the regression coefficient matrix and conditional inverse covariance matrix. All methods are derived in a penalized likelihood framework with hierarchical group penalties. Our theoretical investigation shows that our proposed estimators are consistent and can identify true zero parameters with probability tending to 1 as the sample size goes to infinity. Simulated examples demonstrate that our proposed methods can improve estimation of both regression coefficient matrix and conditional inverse covariance matrix.

In very high dimensional problems, our joint method (DPS) may have numerical difficulty as discussed in Section 2.3. In that case, the proposed plug-in methods are recommended and can often perform better than the DPS method. In certain applications such as our GBM cancer example, a preprocessing step can be first performed before applying the DPS method to reduce dimensions. There is a well-developed literature on preprocessing. For example, see Bair and Tibshirani [21]. With moderate dimensions of predictors and response variables, the joint method can be applied and its performance can be very competitive.

Our current theoretical study is on the case when n goes to infinity. However, for high dimensional cases, it will be also interesting to investigate asymptotic behaviors of our methods when the dimension of predictors, p and the dimension of response variables, m both go to infinity.

Our methods are based on the multivariate Gaussian assumption. Recently, there are some research developments on extending Gaussian graphical models to non-Gaussian cases such as Liu et al. [29] and Cai et al. [30]. Another research direction is to extend our methods to non-Gaussian situations. Further exploration is needed.

Acknowledgments

The authors are partially supported by NSF Grant DMS-0747575 and NIH Grant NIH/NCI R01 CA-149569. The authors are indebted to the editor, the associate editor, and two referees, whose helpful comments and suggestions led to a much improved presentation.

References

1. Breiman, Leo; Friedman, Jerome H. Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society Series B. 1997; 59:3–54.
2. Turlach, Berwin A.; Venables, William N.; Wright, Stephen J. Simulatneous variable selection. Technometrics. 2005; 47:349–363.
3. Yuan, Ming; Ekici, Ali; Lu, Zhaosong; Monteiro, Renato. Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society. Series B. 2007; 69:329–346.
4. Rothman, Adam J.; Levina, Elizaveta; Zhu, Ji. Sparse multiple regression with covariance estimation. Journal of Computational and Graphical Statistics. 2010; 19:947–962.
5. Lee, Wonyul; Liu, Yufeng. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. Journal of Multivariate Analysis. to appear.
6. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
7. Verhaak, Roel GW.; Hoadley, Katherine A.; Purdom, Elizabeth; Wang, Victoria; Qi, Yuan; Wilkerson, Matthew D.; Ryan Miller, C.; Ding, Li; Golub, Todd; Mesirov, Jill P.; Alexe, Gabriele; Lawrence, Michael; OKelly, Michael; Tamayo, Pablo; Weir, Barbara A.; Gabriel, Stacey; Winckler, Wendy; Gupta, Supriya; Jakkula, Lakshmi; Feiler, Heidi S.; Graeme Hodgson, J.; David James, C.; Sarkaria, Jann N.; Brennan, Cameron; Kahn, Ari; Spellman, Paul T.; Wilson, Richard K.; Speed, Terence P.; Gray, Joe W.; Meyerson, Matthew; Getz, Gad; Perou, Charles M.; Neil Hayes, D. TCGA. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]
8. Zhou, Nengfeng; Zhu, Ji. Group variable selection via a hierarchical lasso and its oracle property. Statistics and Its Interface. 2010; 3:557–574.
9. Yuan, Ming; Lin, Yi. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B. 2006; 68:49–67.
10. Zhang, Hao Helen; Liu, Yufeng; Wu, Yichao; Zhu, Ji. Variable selection for the multicategory svm via adaptive sup-norm regularization. Electronic Journal of Statistics. 2008; 2:149–167.
11. Zhao, Peng; Rocha, Guilherme; Yu, Bin. Grouped and hierarchical model selection through composite absolute penalties. The Annals of Statistics. 2009; 37:3468–3497.
12. Yuan, Ming; Lin, Yi. Model selection and estimation in the gaussian graphical model. Biometrika. 2007; 94:19–35.
13. Banerjee, Onureena; El Ghaoui, Laurent; d'Aspremont, Alexandre. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research. 2008; 9:485–516.
14. Friedman, Jerome; Hastie, Trevor; Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]
15. Rothman, Adam J.; Bickel, Peter J.; Levina, Elizaveta; Zhu, Ji. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.
16. Guo, Jian; Levina, Elizabeth; Michailidis, George; Zhu, Ji. Joint estimation of multiple graphical models. Biometrika. 2011; 98:1–15. [PubMed: 23049124]
17. Fan, Jianqing; Li, Runze. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.
18. Zou, Hui. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

19. Knight, Keith; Fu, Wenjiang. Asymptotics for lasso-type estimators. The Annals of Statistics. 2000; 28:1356–1378.
20. Zou, Hui; Li, Runze. One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics. 2008; 36:1509–1533.
21. Bair, Eric; Tibshirani, Robert. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology. 2004; 2:511–522.
22. Zhang, Bin; Horvath, Steve. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4:1–45.
23. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabsi A-L. Hierarchical organization of modularity in metabolic networks. Science. 2002; 297:1151–1155. [PubMed: 12183621]
24. Elsawa, Sherine F.; Almada, Luciana L.; Ziesmer, Steven C.; Novak, Anne J.; Witzig, Thomas E.; Ansell, Stephen M.; Fernandez-Zapico, Martin E. Gli2 transcription factor mediates cytokine cross-talk in the tumor microenvironment. The Journal of Biological Chemistry. 2011; 286:21524–21534. [PubMed: 21454528]
25. Uchida, Hiroyuki; Arita, Kazunori; Yunoue, Shunji; Yonezawa, Hajime; Shinsato, Yoshinari; Kawano, Hiroto; Hirano, Hirofumi; Hanaya, Ryosuke; Tokimura, Hiroshi. Role of sonic hedgehog signaling in migration of cell lines established from cd133-positive malignant glioma cells. J Neurooncol. 2011; 104:697–704. [PubMed: 21380601]
26. Yoon, Hye-Young; Miura, Koichi; Jebb Cuthbert, E.; Davis, Kathryn Kay; Ahvazi, Bijan; Casanova, James E.; Randazzo, Paul A. Arap2 effects on the actin cytoskeleton are dependent on arf6-specific gtpase-activating-protein activity and binding to rhoa-gtp. Journal of Cell Science. 2006; 119:4650–4666. [PubMed: 17077126]
27. Zhou, Xuan; Ren, Yu; Moore, Lynette; Mei, Mei; You, Yongping; Xu, Peng; Wang, Baoli; Wang, Guangxiu; Jia, Zhifan; Pu, Peiyu; Zhang, Wei; Kang, Chunsheng. Downregulation of mir-21 inhibits egfr pathway and suppresses the growth of human glioblastoma cells independent of pten status. Laboratory investigation. 2010; 90:144–155. [PubMed: 20048743]
28. Seike, Masahiro; Goto, Akiteru; Okano, Tetsuya; Bowman, Elise D.; Schetter, Aaron J.; Horikawa, Izumi; Mathe, Ewy A.; Jen, Jin; Yang, Ping; Sugimura, Haruhiko; Gemma, Akihiko; Kudoh, Shoji; Croce, Carlo M.; Harris, Curtis C. Mir-21 is an egfr-regulated anti-apoptotic factor in lung cancer in never-smokers. PNAS. 2009; 106:12085–12090. [PubMed: 19597153]
29. Liu, Han; Lafferty, John; Wasserman, Larry. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research. 2009; 10: 2295–2328.
30. Cai, Tony; Liu, Weidong; Luo, Xi. A constrained l1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011; 106:594–607.

A Proof

A.1 Proof of Theorem 1

A.1.1 Consistency

Let $\boldsymbol{\beta} = (\text{Vec}(\mathbf{B}^{(1)})^T, \ldots, \text{Vec}(\mathbf{B}^{(G)})^T)^T$ and define $Q(\boldsymbol{\beta})$ as

$$Q(\beta)=\sum_{g=1}^{G}\text{tr}\left\{(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})\widehat{\mathbf{C}}^{(g)}(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}+\lambda_1\sum_{j,k}p(\beta_{jk}). \quad (12)$$

To show the results, we use the similar technique in the proof of Theorem 1 in Fan and Li [17]. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\left\{\sup_{||U||=D} Q(\beta^*+\frac{1}{\sqrt{n}}U)>Q(\beta^*)\right\}>1-\delta, \quad (13)$$

where $U = (U^{(1)T}, \ldots, U^{(G)T})^T$ is a $m \times p \times G$-dimensional vector.

Let $\mathbf{y}^{(g)} = \mathrm{Vec}(\mathbf{Y}^{(g)})$, $\mathbf{X}^{m,(g)} = \mathbf{I}_m \otimes \mathbf{X}^{(g)}$ and $\boldsymbol{\beta}^{(g)} = \mathrm{Vec}(\mathbf{B}^{(g)})$; $g = 1, \ldots, G$. Then we can rewrite $Q(\boldsymbol{\beta})$ in (12) as

$$Q(\beta)=\sum_{g=1}^{G}(\mathbf{y}^{(g)}-\mathbf{X}^{m,(g)}\beta^{(g)})^T(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)(\mathbf{y}^{(g)}-\mathbf{X}^{m,(g)}\beta^{(g)})+\lambda_1\sum_{j,k}p(\beta_{jk}). \quad (14)$$

Define $V_n(U)=Q(\beta^*+\frac{1}{\sqrt{n}}U)-Q(\beta^*)$. Using (14), we can show that

$$\begin{aligned}V_n(U)=&\sum_{g=1}^{G}U^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\tfrac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})U^{(g)}-\sum_{g=1}^{G}2\tfrac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)}\\&+\lambda_1\sum_{j,k}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\},\end{aligned} \quad (15)$$

where $\boldsymbol{\varepsilon}^{(g)} = \mathrm{Vec}(\mathbf{e}^{(g)})$; $g = 1, \ldots, G$. Define $\mathcal{I}=\{(j,k)|\beta_{jk}^{*,(g)}\neq 0 \text{ for some } g=1,\ldots,G\}$ for some $g = 1, \ldots, G\}$. Since

$\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}=\left(\sum_{g=1}^{G}\left|\frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}\geq 0,$ for $(j, k) \notin \mathcal{I}$, we have that

$$\begin{aligned}V_n(U)\geq&\sum_{g=1}^{G}U^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\tfrac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})U^{(g)}-\sum_{g=1}^{G}2\tfrac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)}\\&+\lambda_1\sum_{(j,k)\in\mathcal{I}}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}.\end{aligned} \quad (16)$$

For the first term on the right-hand side of (16), note that

$$\sum_{g=1}^{G}U^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\frac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})U^{(g)}=\sum_{g=1}^{G}U^{(g)^T}(\mathbf{C}^{*,(g)}\otimes A^{(g)})U^{(g)}+o_p(1)$$

as $\frac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)} \rightarrow A^{(g)}$ and

$\widehat{\mathbf{C}}^{(g)} \rightarrow_p \mathbf{C}^{*,(g)}$; $g = 1, \ldots, G$.

For the second term on the right-hand side of (16), note that

$$\left|\sum_{g=1}^{G} 2\tfrac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)}\right| \le 2\sum_{g=1}^{G}\left|\left|\tfrac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}\right|\right|\left|\left|U^{(g)}\right|\right|$$
$$\le 2\sum_{g=1}^{G}\left|\left|\tfrac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}\right|\right|\left|\left|U\right|\right|$$
$$= O_p(1)\ \left|\left|U\right|\right|$$

as $\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^{(g)^T}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\mathbf{X}^{m,(g)}\to_d Z$ where Z has multivariate normal distribution. For the third term on the right-hand side of (16), we can show that

$$\lambda_1\sum_{(j,k)\in\mathscr{I}}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}=\lambda_1\sum_{(j,k)\in\mathscr{I}}\sum_{g=1}^{G}\tfrac{1}{\gamma_{jk}}\left\{\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|-\left|\beta_{jk}^{*,(g)}\right|\right\}$$
$$=\tfrac{\lambda_1}{\sqrt{n}}\sum_{(j,k)\in\mathscr{I}}\sum_{g=1}^{G}\tfrac{1}{\gamma_{jk}}\left\{\left|u_{jk}^{(g)}\right|\operatorname{sign}(\beta_{jk}^{*,(g)})+o(1)\right\}=o_p(1),$$

where $\gamma_{jk}=\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}+\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}$.

By combining above statements, we have

$$V_n(U)\ge\sum_{g=1}^{G}U^{(g)^T}(\mathbf{C}^{*,(g)}\otimes A^{(g)})U^{(g)}+O_p(1)\left|\left|U\right|\right|+o_p(1).$$

By choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \| U \| = D\}$ with the probability greater than $1 - \delta$ as $\mathbf{C}^{*,(g)}$ and $A^{(g)}$ are positive-definite. Therefore, (13) holds. This completes the proof of the consistency.

A.1.2 Sparsity

It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any (j, k) such that $\beta_{jk}^{*,(g)}=0$, the partial derivative of Q in (12) with respect to $\beta_{jk}^{(g)}$ at $\widehat{\beta}_{jk}^{(g)}$ has the same sign as $\widehat{\beta}_{jk}^{(g)}$. Let $\boldsymbol{\beta}^{(g)} = \operatorname{Vec}(\mathbf{B}^{(g)})$ and note that

$$\tfrac{\partial}{\partial\beta^{(g)}}(\mathbf{y}^{(g)}-\mathbf{X}^{m,(g)}\beta^{(g)})^T(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{I}_n)\,(\mathbf{y}^{(g)}-\mathbf{X}^{m,(g)}\beta^{(g)})\big|_{\beta^{(g)}=\widehat{\beta}^{(g)}}=(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})\,(\widehat{\beta}^{(g)}-\beta^{*,(g)})-(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{X}^{(g)^T})\boldsymbol{\varepsilon}^{(g)}$$
$$=\sqrt{n}\left\{(\widehat{\mathbf{C}}^{(g)}\otimes\tfrac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})\,\sqrt{n}(\widehat{\beta}^{(g)}-\beta^{*,(g)})-\tfrac{1}{\sqrt{n}}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{X}^{(g)^T})\boldsymbol{\varepsilon}^{(g)}\right\}$$
$$=\sqrt{n}O_p(1)$$

as $(\widehat{\mathbf{C}}^{(g)}\otimes\tfrac{1}{n}\mathbf{X}^{(g)^T}\mathbf{X}^{(g)})\to_p\mathbf{C}^{*,(g)}\otimes A^{(g)}$, $\sqrt{n}(\widehat{\beta}^{(g)}-\beta^{*,(g)})=O_p(1)$ and $\frac{1}{n}(\widehat{\mathbf{C}}^{(g)}\otimes\mathbf{X}^{(g)^T})\boldsymbol{\varepsilon}^{(g)}\to Z$ where Z has multivariate normal distribution. Therefore, the partial derivative of Q can be written as

$$\frac{\partial Q}{\partial \beta_{jk}^{(g)}}\Big|_{\beta_{jk}^{(g)}=\widehat{\beta}_{jk}^{(g)}} = \sqrt{n}O_p(1)+\lambda_1\frac{\text{sign}(\widehat{\beta}_{jk}^{(g)})}{2(\sum_{g=1}^{G}\left|\widehat{\beta}_{jk}^{(g)}\right|)^{1/2}} = \sqrt{n}\left(O_p(1)+\frac{\lambda_1}{n^{1/4}}\frac{\text{sign}(\widehat{\beta}_{jk}^{(g)})}{2(\sum_{g=1}^{G}\left|\sqrt{n}\widehat{\beta}_{jk}^{(g)}\right|)^{1/2}}\right).$$

Since $\frac{\lambda_1}{n^{1/4}} \to \infty$ as $n \to \infty$, the sign of the derivative is completely determined by that of $\widehat{\beta}_{jk}^{(g)}$. This completes the proof of the sparsity.

A.2 Proof of Theorem 2

A.2.1 Consistency

Let $\boldsymbol{c} = (\text{Vec}(\mathbf{C}^{(1)})^T, \ldots, \text{Vec}(\mathbf{C}^{(G)})^T)^T$ and define $Q(\boldsymbol{c})$ as

$$Q(\mathbf{c}) = \sum_{g=1}^{G}\left\{-n\log\det(\mathbf{C}^{(g)})+n\text{tr}(\mathbf{S}^{(g)}\mathbf{C}^{(g)})\right\}+\lambda_2\sum_{s\neq t}\left(\sum_{g=1}^{G}\left|c_{st}^{(g)}\right|\right)^{1/2} \quad (17)$$

To show the results, we use the similar technique in the proof of Theorem 1. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\left\{\sup_{||U||=D} Q(\boldsymbol{c}^*+\frac{1}{\sqrt{n}}U) > Q(\boldsymbol{c}^*)\right\} > 1-\delta, \quad (18)$$

where $U = (\text{Vec}(U^{(1)})^T, \ldots, \text{Vec}(U^{(G)})^T)^T$ is a $m \times m \times G$-dimensional vector.

Using (17), define $V_n(U)$ as

$$V_n(U) = Q(\boldsymbol{c}^*+\tfrac{1}{\sqrt{n}}U)-Q(\boldsymbol{c}^*)$$
$$= \sum_{g=1}^{G}\left\{-n\log\det((\mathbf{C}^{*,(g)}+\tfrac{U^{(g)}}{\sqrt{n}})(\mathbf{C}^{*,(g)})^{-1})+n\text{tr}(\tfrac{U^{(g)}\mathbf{S}^{(g)}}{\sqrt{n}})\right\}+\lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|c_{st}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|c_{jk}^{*,(g)}\right|\right)^{1/2}\right\}.$$

Using the similar argument as in the proof of Lemma 2 in Lee and Liu [5], it can be shown that

$$V_n(U) = \sum_{g=1}^{G}\text{tr}(U^{(g)}\sum^{(g)}U^{(g)}\sum^{(g)})+\sum_{g=1}^{G}\text{tr}[U^{(g)}\sqrt{n}(\mathbf{S}^{(g)}-\sum^{(g)})]$$
$$+\lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|c_{st}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|c_{jk}^{*,(g)}\right|\right)^{1/2}\right\}+o(1). \quad (19)$$

For the second term on the right-hand side of (19), by using the similar argument as in the proof of Theorem 2 in Lee and Liu [5], it can be shown that

$$\sum_{g=1}^{G}\mathrm{tr}[U^{(g)}\sqrt{n}(\mathbf{S}^{(g)}-\sum\nolimits^{(g)})]=\sum_{g=1}^{G}\mathrm{tr}[U^{(g)}\sqrt{n}(\mathbf{S}^{*,(g)}-\sum\nolimits^{(g)})]+o_p(1)$$

where $\mathbf{S}^{*,(g)}=\frac{1}{n}(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{*,(g)})\,(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{*,(g)})^{T}$. Note that $\sqrt{n}(\mathbf{S}^{*,(g)}-\sum^{(g)})$ converges in distribution to multivariate normal distribution by the central limit theorem.

For the third term on the right-hand side of (19), by using the similar argument as in the proof of Theorem 1, it can be shown that

$$\lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|c_{st}^{*,(g)}+\frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|c_{jk}^{*,(g)}\right|\right)^{1/2}\right\}=o_p(1).$$

By combining the above statements, we can conclude that the first term on the right-hand side of (19) dominates the other terms. Therefore, by choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \| U \| = D\}$ with the probability greater than $1 - \delta$. This completes the proof of the consistency.

A.2.2 Sparsity

Similar to the proof of the sparsity in Theorem 1, it is sufficient to show that with probability tending to 1 as $n \to \infty$, for any (s, t) such that $c_{st}^{*,(g)}=0$, the partial derivative of Q in (17) with respect to $c_{st}^{(g)}$ at $\hat{c}_{st}^{(g)}$ has the same sign as $\hat{c}_{st}^{(g)}$. Note that

$$\frac{\partial Q}{\partial c_{st}^{(g)}}\Big|_{c_{st}^{(g)}=\hat{c}_{st}^{(g)}}=n\,(\mathbf{s}_{st}^{(g)}-\hat{\sigma}_{st}^{(g)})+\lambda_2\frac{\mathrm{sign}(\hat{c}_{st}^{(g)})}{2(\sum_{g=1}^{G}\left|\hat{c}_{st}^{(g)}\right|)^{1/2}}$$

where $\mathbf{S}^{(g)}=(\mathbf{s}_{st}^{(g)})$ and $(\hat{\mathbf{C}}^{(g)})^{-1}=(\hat{\sigma}_{st}^{(g)})$. By using the argument in the proof of Theorem 2 in Guo et al. [16], one can show that $(\mathbf{s}_{st}^{(g)}-\hat{\sigma}_{st}^{(g)})=O_p(1/\sqrt{n})$. Therefore, we have

$$\frac{\partial Q}{\partial c_{st}^{(g)}}\Big|_{c_{st}^{(g)}=\hat{c}_{st}^{(g)}}=\sqrt{n}\left(O_p(1)+\frac{\lambda_2}{n^{1/4}}\frac{\mathrm{sign}(\hat{c}_{st}^{(g)})}{2(\sum_{g=1}^{G}\left|\sqrt{n}\hat{c}_{st}^{(g)}\right|)^{1/2}}\right).$$

Since $\frac{\lambda_2}{n^{1/4}} \to \infty$ as $n \to \infty$, the sign of the derivative is completely determined by that of $\hat{c}_{st}^{(g)}$. This completes the proof of the sparsity.

A.3 Proof of Theorem 3

A.3.1 Consistency

Define $Q(\boldsymbol{\beta}, \boldsymbol{c})$ as

$$Q(\beta,\boldsymbol{c})=\sum_{g=1}^{G}\left\{-l_g(\beta,\boldsymbol{c})+\lambda_1\sum_{jk}\left(\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|\right)^{1/2}+\lambda_2\sum_{s\neq t}\left(\sum_{g=1}^{G}\left|c_{st}^{(g)}\right|\right)^{1/2}\right\}, \quad (20)$$

where $l_g(\boldsymbol{\beta}, \boldsymbol{c}) = n \log\det(\mathbf{C}^{(g)}) - \mathrm{tr}\,(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T$.

To show the results, we use the similar technique in the proof of Theorem 1. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\left\{\sup_{\left||U|\right|=D} Q(\beta^*+\frac{1}{\sqrt{n}}U_1,\boldsymbol{c}^*+\frac{1}{\sqrt{n}}U_2)>Q(\beta^*,\boldsymbol{c}^*)\right\}>1-\delta, \quad (21)$$

where $U=(U_1^T,U_2^T)^T, U_1=(\mathrm{Vec}(U_1^{(1)})^T,\ldots,\mathrm{Vec}(U_1^{(G)})^T)$ and $U_2=(\mathrm{Vec}(U_2^{(1)})^T,\ldots,\mathrm{Vec}(U_2^{(G)})^T)$

Using (20), define $V_n(U)=Q(\beta^*+\frac{1}{\sqrt{n}}U_1,\boldsymbol{c}^*+\frac{1}{\sqrt{n}}U_2)-Q(\beta^*,\boldsymbol{c}^*)$. It can be shown that

$$\begin{aligned}V_n(U)=&\sum_{g=1}^{G}\left\{-n\log\det((\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\mathbf{C}^{*,(g)})^{-1})+n\mathrm{tr}(\frac{U_2^{(g)}\mathbf{S}^{*,(g)}}{\sqrt{n}})\right\}\\&+\sum_{g=1}^{G}\left\{\mathrm{tr}\left[(\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})^T(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})\right]\right\}\\&-2\sum_{g=1}^{G}\left\{\mathrm{tr}\left[(\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{*,(g)})^T(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})\right]\right\}\\&+\lambda_1\sum_{j,k}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\frac{1}{\sqrt{n}}u_{1,jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}\\&+\lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|c_{st}^{*,(g)}+\frac{1}{\sqrt{n}}u_{2,jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|c_{jk}^{*,(g)}\right|\right)^{1/2}\right\}.\end{aligned} \quad (22)$$

For the first term on the right-hand side of (22), it has been shown in Theorem 2 that

$$\sum_{g=1}^{G}\left\{-n\log\det((\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\mathbf{C}^{*,(g)})^{-1})+n\mathrm{tr}(\frac{U_2^{(g)}\mathbf{S}^{*,(g)}}{\sqrt{n}})\right\}=\sum_{g=1}^{G}\mathrm{tr}(U_2^{(g)}\sum^{(g)}U_2^{(g)}\sum^{(g)})+O_p(1)$$

For the second term and the third term on the right-hand side of (22), by using the similar argument in the proof of Lemma 3 in Lee and Liu [5], it can be shown that

$$\sum_{g=1}^{G}\left\{\mathrm{tr}\left[(\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})^T(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})\right]\right\}\sum_{g=1}^{G}U_1^{(g)^T}(\mathbf{C}^{*,(g)}\otimes A^{(g)})U_1^{(g)}+o_p(1)$$

and

$$\sum_{g=1}^{G}\left\{\mathrm{tr}\left[(\mathbf{C}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})\,(\boldsymbol{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{*,(g)})^T(\frac{\mathbf{X}^{(g)}\mathbf{U}_1^{(g)}}{\sqrt{n}})\right]\right\}=O_p(1).$$

For the fourth and fifth term on the right-hand side of (22), it has been shown in Theorems 1 and 2 that

$$\lambda_1\sum_{j,k}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{1,jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}=o_p(1),$$
$$\lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|c_{st}^{*,(g)}+\tfrac{1}{\sqrt{n}}u_{2,jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|c_{jk}^{*,(g)}\right|\right)^{1/2}\right\}=o_p(1).$$

By combining the above statements, we can conclude that the right-hand side of (22) is dominated by $\sum_{g=1}^{G}\mathrm{tr}(U_2^{(g)}\sum^{(g)}U_2^{(g)}\sum^{(g)})$ and $\sum_{g=1}^{G}U_1^{(g)^T}(\mathbf{C}^{*,(g)}\otimes A^{(g)})U_1^{(g)}$. Therefore, by choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \| U \| = D\}$ with the probability greater than $1 - \delta$. This completes the proof of the consistency.

A.3.2 Sparsity

Note that $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{c}})$ is a $\sqrt{n}$-consistent local minimizer of $Q(\boldsymbol{\beta}, \boldsymbol{c})$ defined in (20). As $\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}\, Q(\boldsymbol{\beta}, \hat{\boldsymbol{c}})$ and $\hat{\boldsymbol{c}}$ is $\sqrt{n}$-consistent, the sparsity of $\hat{\boldsymbol{\beta}}$ holds by Theorem 1. Similarly, since $\hat{\boldsymbol{c}} = \mathrm{argmin}_{\boldsymbol{c}}\, Q(\hat{\boldsymbol{\beta}}, \boldsymbol{c})$ and $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, the sparsity of $\hat{\boldsymbol{c}}$ holds by Theorem 2. These complete the proof of this theorem.

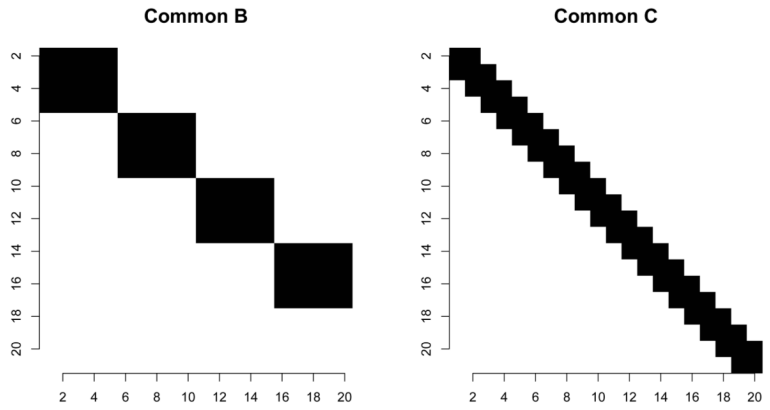

Figure 1.
Regression Parameter Structure and Inverse Covariance Structure that are common in all groups. Non-zero entries are colored as black and zero entries are colored as white.

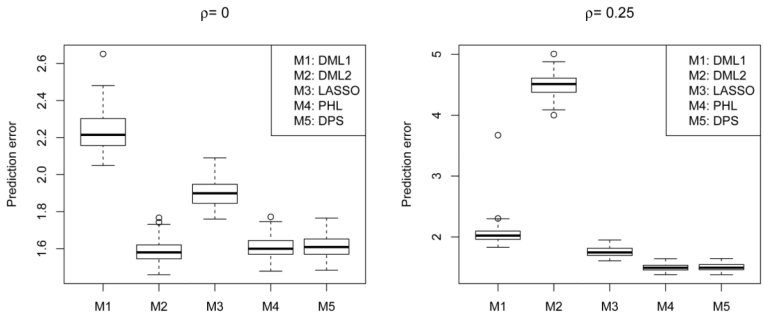

Figure 2.
Boxplots of prediction errors of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

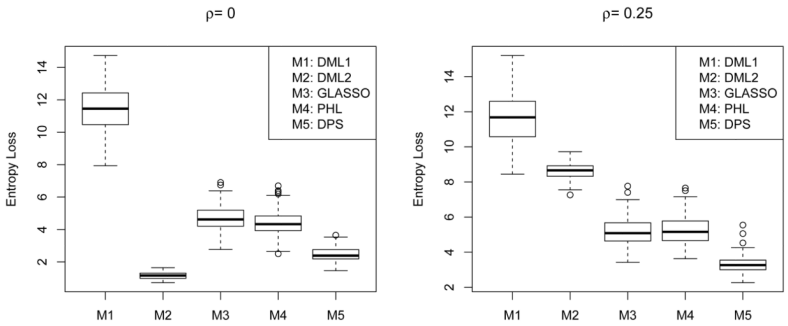

Figure 3.
Boxplots of entropy losses of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

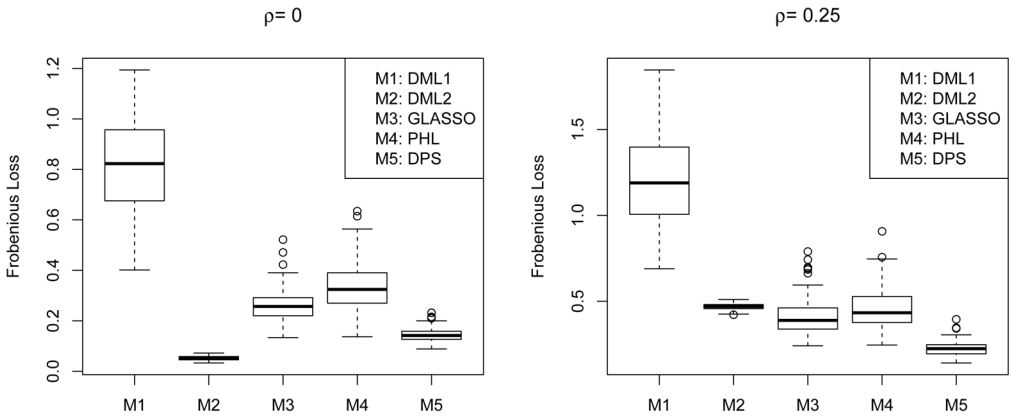

Figure 4.

Boxplots of Frobenius losses of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

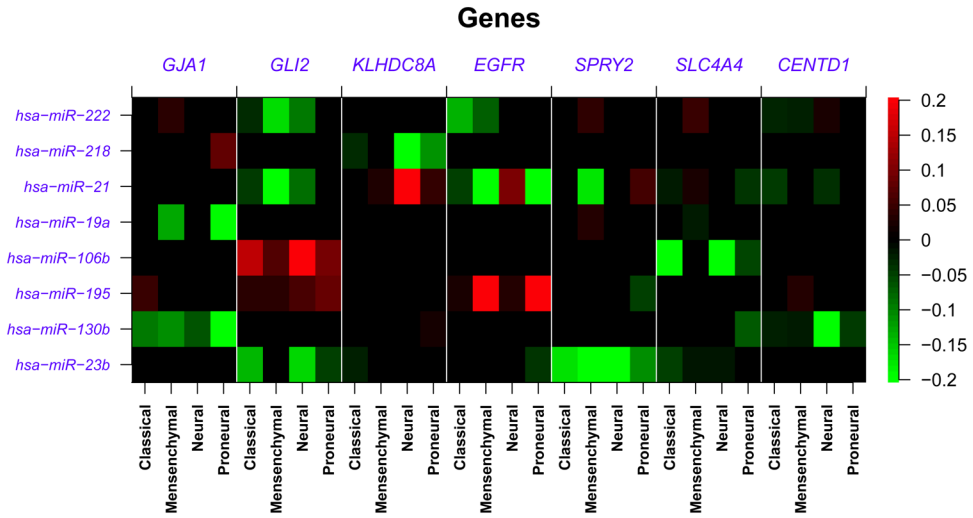

Figure 5.
Heatmap of averaged estimated regression coefficients of several micro-RNAs for some selected genes. The DPS estimates are used to generate the heatmap.

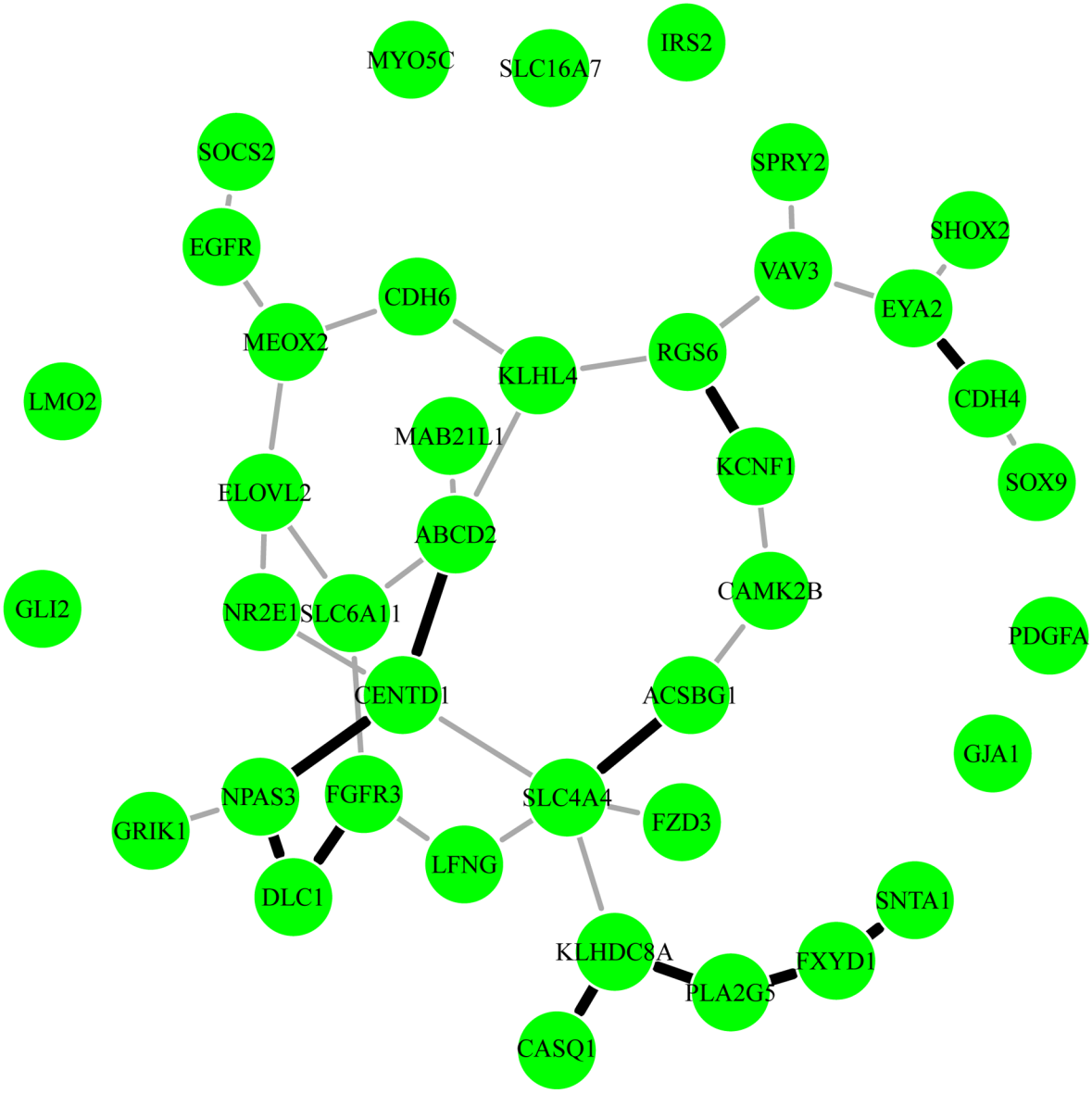

Figure 6.
A graphical model of gene expressions based on the estimated inverse covariance matrix. Black lines are common edges across all subgroups. Grey lines are unique edges to some subgroups. The DPS estimates are used to generate the network.

Table 1

Average prediction error, entropy loss, and Frobenius loss based on 100 replications (The numbers in parentheses are standard errors)

	ρ	**M1: DML1**	**M2: DML2**	**M3: LASSO**	**M4: PHL**	**M5: DPS**
Prediction Error	0	2.23(0.011)	1.59(0.006)	1.90(0.008)	1.61(0.006)	1.61(0.006)
	0.25	2.05(0.021)	4.51(0.018)	1.76(0.009)	1.50(0.007)	1.50(0.007)
	ρ	M1: DML1	M2: DML2	M3: GLASSO	**M4: PHGL**	**M5: DPS**
Entropy Loss	0	11.52(0.153)	1.17(0.020)	4.69(0.077)	4.40(0.079)	2.47(0.043)
	0.25	11.58(0.149)	8.62(0.046)	5.22(0.058)	5.27(0.087)	3.31(0.051)
Frobenius Loss	0	0.82(0.019)	0.05(0.001)	0.36(0.006)	0.34(0.010)	0.15(0.003)
	0.25	1.20(0.027)	0.47(0.002)	0.42(0.012)	0.46(0.012)	0.22(0.004)

Table 2

Averages of relative computational time of M4 and M5 compared with M3 based on 100 replications (The numbers in parentheses are standard errors). For example, when $\rho = 0$, the computational time of M4 is 3.92 times of that for M3.

		M3	M4: PHL	M5: DPS
Simulated examples	$\rho = 0$	1	3.92(0.05)	38.40(0.35)
	$\rho = 0.25$	1	3.44(0.04)	30.09(0.35)

Table 3

Averages of PE based on 100 replications (The numbers in parentheses are standard errors)

	DML	LASSO	PHL	DPS
PE	1.373(0.004)	1.025(0.003)	1.050(0.004)	1.065(0.004)