New Statistical Learning Methods for Multiple High Dimensional Datasets

Wonyul Lee

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2013

Approved by:

Dr. Yufeng Liu

Dr. Jan Hannig

Dr. D. Neil Hayes

Dr. J. S. Marron

Dr. Wei Sun

Abstract

WONYUL LEE: New Statistical Learning Methods for Multiple High Dimensional Datasets
(Under the direction of Dr. Yufeng Liu)

In this dissertation, we design several statistical learning methods for analyzing multiple high-dimensional datasets. Our focus is on multiple response regression and inverse covariance matrix estimation.

Multivariate regression is a common statistical tool for practical problems. Many multivariate regression techniques are designed for univariate response cases. For problems with multiple response variables available, one common approach is to apply the univariate response regression technique separately on each response variable. Although it is simple and popular, the univariate response approach ignores the joint information among response variables. We propose several methods for utilizing joint information among response variables in a penalized likelihood framework. The proposed methods provide sparse estimators for the conditional inverse covariance matrix of response vector given explanatory variables as well as sparse estimators of regression coefficient matrix.

Estimation of inverse covariance matrices is important in various areas of statistical analysis. The task of estimating multiple inverse covariance matrices sharing some common structure is considered. In this case, estimating each matrix separately can be suboptimal as it ignores potential common structure. We propose a new approach to parameterize each inverse covariance matrix as a sum of common and unique components and jointly estimate multiple inverse covariance matrices in a constrained L_1 minimization framework.

Theoretical properties of the new methods are explored. Simulated examples and applications to a glioblastoma multiforme cancer data demonstrate competitive performance of the proposed methods.

Acknowledgments

I wish to deeply thank my advisor, Professor Yufeng Liu, for his guidance, encouragement, and support during my PhD research. His insightful advice enabled me to enjoy my research and complete my dissertation work successfully. I could not imagine completing my PhD degree without his inspiring suggestions and encouragement. Beyond my research work, I owe many thanks to him for his general guidance about my future academic path.

I would also like to express my sincere appreciation to committee members, J.S. Marron, Jan Hannig, Wei Sun , D. Neil Hayes for their valuable comments and suggestions on my dissertation.

Finally, my deepest gratitude goes to my wife, Hana, for her love and support. She always believes in my potential, which greatly encourages me to keep advancing. Without her belief and encouragement, I would not have completed my PhD study. For my entire life, I will always love and respect you, Hana.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Some Background on Regression

Regression is one of the most fundamental tools in statistics. It helps to build a model to characterize the relationship between predictors and response. A regression model can be very useful for both prediction and interpretation. In this section, we briefly review some regression techniques. In Section 1.1.1, we focus on univariate response regression and in Section 1.1.2, we discuss multivariate response regression techniques. For the simplicity of notations, we assume that all response variables are centered so that regression models do not include intercept terms.

1.1.1 Univariate Response Regression

In the statistical learning literature, multivariate regression is a common and popular technique that builds a model to predict a response variable given a set of predictor variables. In particular, we have a training sample $\{(\mathbf{x}_i, y_i)\}_{i=1,...,n}$, where $\mathbf{x}_i \in \mathbf{R}^p$ is a p-dimensional vector of predictors and y_i is a centered response variable, i.e., $\sum_{i=1}^n y_i = 0$. The multivariate regression model has the form

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \quad \text{for } i = 1, ..., n, \tag{1.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients and ϵ_i denotes the random error term.

The ordinary least squares (OLS) method is one of the most common multivariate

regression techniques to estimate the regression coefficients. Even though it has been widely used in the literature, it can have difficulty when the dimension of predictors p is large. In particular, the problem of overfitting may arise and consequently decrease the prediction accuracy. In addition to that, the resulting OLS model often keeps all variables in the model. It can be undesirable when only a small subset of predictors truly influence the response variable. Therefore, variable selection is an important issue in the multivariate regression problem. Accurate variable selection can not only improve prediction accuracy, but also provide better interpretability of the model.

Many variable selection techniques are available in the literature. Traditionally, the approach of subset selection has been widely used to select important variables. In this approach, we select the subset of predictors first and then fit the regression model on the selected predictor set. There are many different subset selection techniques, for example, forward stepwise selection, backward stepwise selection, and the combination of forward and backward stepwise selection. Theses techniques are simple to implement. However, they can be unstable because the procedure is not continuous [Breiman, 1996].

Recently, a large number of methods based on the regularization framework have been proposed. Some well-known methods in this group include the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani [1996], the nonnegative garrote proposed by Breiman [1995], and the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li [2001]. In the regularization framework, the objective function for us to optimize is in the form of *loss+penalty*. The loss term measures goodness of fit for our model and the penalty term measures complexity of models which helps to control overfitting. In particular, the optimization problem for the penalized least squares can be written as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda p(\boldsymbol{\beta}), \tag{1.2}$$

where λ is a tuning parameter to balance the two terms. In practice, λ needs to be chosen carefully. If we choose a sparse penalty term for $p(\boldsymbol{\beta})$, as a consequence, the estimated coefficients are shrunk toward 0 with some of them exactly being 0. Those predictors with nonzero coefficients remain in the model. Thus, such sparse regularization techniques can perform variable selection and model estimation simultaneously. For example, LASSO uses

the L_1 norm of $\boldsymbol{\beta}$, $\sum_{j=1}^{p} |\beta_j|$, as $p(\boldsymbol{\beta})$.

1.1.2 Multivariate Response Regression

In Section 1.1.1, the focus has been on multivariate regression model with one response variable. However, in many applications, one may have multiple response variables with the same set of predictor variables. In that case, multiple response regression is a useful regression technique to tackle this problem. In particular, with an m-dimensional vector of response variables, $\mathbf{y}_i = (y_{i1}, ..., y_{im})$, the multiple response regression model can be formulated by generalizing (1.1) as follows,

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\epsilon}_i \quad \text{for } i = 1, ..., n,$$

where $\mathbf{B}$ is a $p \times m$ matrix of regression coefficients and an $\boldsymbol{\epsilon}_i$ denotes an m-dimensional error vector.

The standard approach to estimate the regression parameter matrix $\mathbf{B}$ is to regress each response variable separately on the same set of predictor variables. All marginal univariate regression procedures discussed in Section 1.1.1 can be applied to each response. For example, one can apply the OLS method to each response separately by solving

$$\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} (y_{ij} - \boldsymbol{\beta}_j^T \mathbf{x}_i)^2, \quad \text{for } j = 1, ..., m, \tag{1.3}$$

where $\boldsymbol{\beta}_j$ is the j-th column of $\mathbf{B}$. By using simple linear algebra, it can be shown that the optimization problem of (1.3) is equivalent to the following optimization problem

$$\min_{\mathbf{B}} \operatorname{tr}[(\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB})], \tag{1.4}$$

where $\mathbf{Y}$ is the $n \times m$ response matrix, $\mathbf{X}$ is the $n \times p$ predictor matrix. Even though it is simple to implement, this separate approach may not be optimal since they do not utilize the joint information among response variables.

To utilize the correlation information among response variables, Breiman and Friedman [1997] proposed a method, called Curd and Whey ($C\&W$). The $C\&W$ method predicts the multiple responses with an optimal linear combination of the OLS estimators. In particular,

the $C\&W$ procedure starts with fitting m separate ordinary least squares models. With the resulting predictor $\hat{\mathbf{y}}^{OLS}$ available, the $C\&W$ method tries to find out another predictor $\tilde{\mathbf{y}} = \mathbf{W}\hat{\mathbf{y}}^{OLS}$ with an optimal $m \times m$ matrix $\mathbf{W}$ so that

$$\mathbf{E}\{(y_j - (\mathbf{W}\hat{\mathbf{y}}^{OLS})_j)\}^2 \leq \mathbf{E}\{(y_j - (\hat{\mathbf{y}}^{OLS})_j)\}^2, \quad j = 1, ..., m.$$

In other words, $\mathbf{W}$ reduces mean-squared prediction error for each response. They showed that $\mathbf{W}$ can be obtained by canonical analysis. In particular, canonical analysis seeks pairs of linear combination such that

$$\begin{aligned} &(\mathbf{t}_k, \mathbf{v}_k) = \arg\max_{\mathbf{t},\mathbf{v}} \mathrm{Corr}(\mathbf{t}^T\mathbf{y}, \mathbf{v}^T\mathbf{x}) \\ &\text{subject to} \quad \mathrm{Corr}(\mathbf{t}^T\mathbf{y}, \mathbf{t}_l^T\mathbf{y}) = 0, \mathrm{Corr}(\mathbf{v}^T\mathbf{x}, \mathbf{v}_l^T\mathbf{x}) = 0, l = 1, ..., k-1, \end{aligned} \tag{1.5}$$

where $k = 1, ..., \min(p, m)$ and $\mathrm{Corr}(a, b)$ is the correlation between a and b. Then $\mathbf{W}$ is given by $\mathbf{W} = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}$, where $\mathbf{T}$ is the $m \times m$ matrix whose k-th row is $\mathbf{t}_k$ and $\mathbf{D}$ is a diagonal matrix. The i-th diagonal entry of $\mathbf{D}$ is $d_i = \rho_i^2/[\rho_i^2 + r(1 - \rho_i^2)]$, where $\rho_i = \mathrm{Corr}(\mathbf{t}_i^T\mathbf{y}, \mathbf{v}_i^T\mathbf{x})$ and $r = p/n$. As diagonal elements in $\mathbf{D}$ are less than or equal to 1, the $C\&W$ method achieves multivariate shrinkage after transforming $\hat{\mathbf{y}}^{OSL}$. They showed that their method can outperform separate univariate regression approaches when there are correlations among the response variables.

Some other approaches to tackle multiple response regression problem have been proposed in the regularization framework [Yuan et al., 2007; Turlach, Venables and Wright, 2005]. These approaches impose a constraint on the parameters to stabilize the estimators. In particular, they solve the following optimization problem

$$\min_{\mathbf{B}} \mathrm{tr}[(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})] \quad \text{subject to:} \quad J(\mathbf{B}) \leq t,$$

where $J(\mathbf{B})$ is a constraint function and t is a tuning parameter. Without any constraint, the objective function is identical to (1.4) which is the objective function of separate OLS approach. However, by imposing a constraint, we achieve shrinkage in the resulting estimator. In particular, Yuan et al. [2007] proposed a method called factor estimation and

selection. To encourage sparsity among singular values of the regression parameter matrix, they employed $J(\mathbf{B}) = \sum_i^{\min(p,m)} \sigma_i(\mathbf{B})$, where $\sigma_i(\mathbf{B})$ is the ith singular value of $\mathbf{B}$. As a result, their method achieves dimension reduction in $\mathbf{B}$. Turlach, Venables and Wright [2005] proposed another constraint function, $J(\mathbf{B}) = \sum_{j=1}^{p} \max(|\beta_{j1}|, ..., |\beta_{jm}|)$. By imposing the max-$L_1$ penalty, they select a common subset of explanatory variables which can be used as predictors for all response variables.

1.2 Background on Gaussian Graphical Model

Gaussian graphical models explore conditional dependence structure among variables under the multivariate Gaussian distributional assumption. In particular, let $\mathbf{x}$ be a p-dimensional vector following a multivariate normal distribution $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is an unknown p-dimensional mean vector and $\boldsymbol{\Sigma}$ is a nonsingular covariance matrix. The conditional dependence structure can be determined from the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ij})_{1 \le i,j \le p}$. In particular, an off-diagonal element ω_{ij} in $\boldsymbol{\Omega}$ is proportional to the conditional correlation between variable i and j given the other variables. In other words, ω_{ij} is zero if and only if variable i and j are conditionally independent given the other variables. Therefore, in a Gaussian graphical model, one of the main interests is to identify zero entries in the precision matrix. The nonzero entries in $\boldsymbol{\Omega}$ correspond to conditionally correlated pairs of variables given other variables.

For illustration, let $\mathbf{x}$ be a 5-dimensional vector following a multivariate normal distribution $\mathbf{N}(0, \boldsymbol{\Sigma}_5)$ with

$$\boldsymbol{\Omega}_5 = \boldsymbol{\Sigma}_5^{-1} = \begin{bmatrix} 3 & 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix}.$$

The corresponding undirected graph is depicted in Figure 1.1. Each variable in $\mathbf{x}$ represents a node in the graph. If ω_{ij} is nonzero, then the nodes x_i and x_j are connected as they are conditionally correlated. For example, there is an edge between x_1 and x_3 as ω_{13} is not zero. On the other hand, x_1 and x_2 are not connected in the graph since ω_{12} is zero.

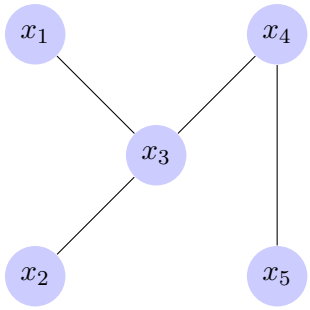

Figure 1.1: The undirected graph of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$.

A standard approach to perform model selection in Gaussian graphical models is the backward stepwise selection method. The procedure starts with a fully connected graph. In each step, deletion of the least significant edge is performed based on hypothesis testing at some prespecified significance level α. The process continues until all remaining edges are significant and cannot be deleted. Once the model selection procedure is finished, then parameter estimation is performed based on the selected model. However, this procedure does not take multiple comparisons into account [Edwards, 2000].

In recent years, various penalized maximum likelihood methods have been proposed for the estimation of sparse Gaussian graphical models [Yuan and Lin, 2007; Banerjee, Ghaoui and d'Aspremont, 2008; Friedman, Hastie and Tibshirani, 2008; Rothman et al., 2008; Lam and Fan, 2009; Fan, Feng and Wu, 2009]. These approaches produce a sparse estimator of the precision matrix by maximizing the penalized Gaussian likelihood with sparse penalties such as the L_1 penalty and the smoothly clipped absolute deviation penalty [Fan and Li, 2001]. In particular, we solve the following optimization problem

$$\min_{\mathbf{\Omega}} \left\{ -\log\det(\mathbf{\Omega}) + \mathrm{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}) + \lambda \sum_{i \neq j} p(\omega_{ij}) \right\}, \tag{1.6}$$

where $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix of variables, λ is a tuning parameter, and $\sum_{i \neq j} p(\omega_{ij})$ is a sparse penalty function. Note that the first two terms in (1.6), $-\log\det(\mathbf{\Omega}) + \mathrm{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}})$, correspond to the negative log Gaussian likelihood up to a constant not depending on $\mathbf{\Omega}$. By imposing a sparse penalty on the off-diagonal elements of the precision matrix, they encourage sparsity among off-diagonal entries in the estimator of $\mathbf{\Omega}$.

Instead of using likelihood approaches, several techniques take advantage of the connection between linear regression and the entries of the precision matrix [Meinshausen and

Buhlmann, 2006; Peng et al., 2009; Yuan, 2010]. In particular, these approaches convert the estimation problem of the precision matrix into relevant regression problems and solve them with sparse regression techniques accordingly. One advantage of these approaches is that they can handle a wide range of distributions including the Gaussian case.

1.3 New Contributions and Outline

As discussed in Sections 1.1 and 1.2, there are many existing papers in the literature focusing on

(1) the case of multivariate regression problems with a single response variable;

(2) estimation of the inverse covariance matrix of a multivariate Gaussian data set with the applications to the Gaussian graphical model.

In Chapters 2-3, our focus is to combine both goals through the setting of multiple response multivariate regression. Our proposed methodology provides a good insight on the effect of the joint estimation of regression parameters and the inverse of residue covariance matrix. In Chapter 4, we focus on the joint estimation of multiple inverse covariance matrices.

- In Chapter 2, we propose three new methods to estimate the regression parameter matrix and the conditional inverse covariance matrix in the penalized Gaussian maximum likelihood framework [Lee and Liu, 2012]. Our methods and the corresponding theoretical developments show that compared to separate modeling, simultaneous modeling of the multiple response variables can provide more accurate estimation of both regression parameters and the inverse covariance matrix.

- In Chapter 3, we consider the data coming from a mixture of several Gaussian distributions. In particular, we extend the methods proposed in Chapter 2 with the hierarchical group penalty to address the mixture structure [Lee et al., 2012]. With the proposed methods, multiple groups from different Gaussian distributions can be modeled jointly. In these approaches, we allow the common structure across different groups and at the same time can estimate unique structure to each group. We establish some asymptotic properties of the methods. Both simulated examples and

an application to a glioblastoma cancer dataset are presented to demonstrate the performance of our methods.

- In Chapter 4, we consider estimation of multiple inverse covariance matrices sharing some common structure. To estimate potential common structure more efficiently, we propose a new approach to parameterize each precision matrix as a sum of common and unique components and estimate multiple precision matrices in a constrained L_1 minimization framework [Lee and Liu, 2013]. Some theoretical properties of the method are derived in the high dimensional setting. Numerical examples are presented as well to illustrate the advantage of our proposed method.

Chapter 2

Multiple Response Regression with Sparse Inverse Covariance Estimation

2.1 Introduction

With multiple response variables available, the standard approach to model them is to regress each response variable separately on the same set of explanatory variables. All marginal univariate regression procedures discussed in Section 1.1.1 can be applied to each response variable. However, this approach may not be optimal since they do not utilize the information among response variables. As stated in Section 1.1.2, Breiman and Friedman [1997] proposed the $C\&W$ method that uses the relationship among response variables to improve predictive accuracy. They showed that their method can outperform separate univariate regression approaches when there are correlations among the response variables. However, their method did not address the topic of variable selection. Recently, Yuan et al. [2007] proposed a method based on dimension reduction as stated in Section 1.1.2. Their idea is to obtain dimension reduction by encouraging sparsity among singular values of the parameter matrix. However, their approach focuses on dimension reduction rather than variable selection. Thus, it does not give a subset of explanatory variables for each response. Variable selection can be a very important issue when the number of explanatory variables is large or when explanatory variables are highly correlated. To relate with variable selection, Turlach, Venables and Wright [2005] proposed a penalized method using the max-L_1 penalty to select a common subset of explanatory variables for multiple response regression. However, this assumption may be too strong when each response has different sets of explanatory variables. A similar technique was proposed by Zhang et al. [2008] for

multicategory support vector machines.

In this chapter, we propose three approaches to tackle the multiple response regression problem via utilizing the joint information among multiple response variables. To handle the problem, we need to estimate two parameter matrices, the regression parameter matrix $\mathbf{B}$ and the conditional inverse covariance matrix of response variable $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. The first two approaches are plug-in methods, i.e., plugging in an estimator of one parameter matrix to solve the other one. The third approach tries to jointly estimate both parameter matrices. In particular, for the first approach, we plug in a reasonable estimator of $\mathbf{\Omega}$ to estimate the regression parameter matrix $\mathbf{B}$. For the second approach, we estimate $\mathbf{\Omega}$ instead after plugging in a good estimator of $\mathbf{B}$. The last approach simultaneously estimates the regression parameter matrix $\mathbf{B}$, and the inverse covariance matrix $\mathbf{\Omega}$. These methods are penalized log-likelihood approaches with the multivariate Gaussian assumption. The first proposed method maximizes a sparse penalized log-likelihood using a previously estimated inverse covariance matrix $\hat{\mathbf{\Omega}}$. Similarly, the second proposed method maximizes a sparse penalized log-likelihood using a previously estimated regression parameter matrix $\hat{\mathbf{B}}$. The last proposed method simultaneously estimates regression parameters and the inverse covariance matrix by maximizing a doubly penalized joint likelihood function. These methods involve two penalty terms: the weighted L_1 penalty on the inverse covariance matrix $\mathbf{\Omega}$ and the weighted L_1 penalty on the regression parameter matrix $\mathbf{B}$. Note that the joint approach was also considered recently in Rothman, Levina and Zhu [2010] with unweighted L_1 penalty terms. Our framework allows flexible weights on the penalty terms and it is more general. Besides the regression predictive accuracy, we also study the performance of the estimation of $\mathbf{\Omega}$. Furthermore, we establish theoretical properties of all three methods. To handle the computational difficulty of high dimensional problems, we also suggest some prescreening procedure to eliminate noise variables before further estimation.

In the following sections, we describe the new proposed methods in more details with theoretical justification and numerical examples. In Section 2.2, we introduce our proposed methodology. Section 2.3 explores theoretical properties of our proposed methods. Section 2.4 develops coordinate descent computational algorithms to obtain solutions for proposed methods. A prescreening step is suggested for the joint method to speed up the computation. Simulated examples are presented in Section 2.5 to demonstrate performance of our methods

and Section 2.6 provides a glioblastoma cancer data example. Section 2.7 provides some discussions. The proofs of the theorems are provided in Section 2.8.

2.2 Methodology

Consider the regression problem of p covariates and m response variables. Suppose the data contain n observations. Let $\mathbf{y}_i = (y_{i1}, ..., y_{im})^T; i = 1, ..., n$, be m-dimensional responses and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^T$ be the $n \times m$ response matrix. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T; i = 1, ..., n$, be p-dimensional predictors and $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$ be the $n \times p$ design matrix. For simplicity of the notation, let $\mathbf{y}^k = (y_{1k}, ..., y_{nk})^T$ be the k-th response vector $(k = 1, ..., m)$ and $\mathbf{x}^j = (x_{1j}, ..., x_{nj})^T$ be the j-th predictor $(j = 1, ..., p)$. Consider the following model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}, \quad \text{where} \quad \mathbf{e} = [\epsilon_1, ..., \epsilon_n]^T,$$

where $\mathbf{B} = \{\beta_{jk}\}; j = 1, ..., p, k = 1, ..., m$, is an unknown $p \times m$ parameter matrix. The errors $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{im})^T; i = 1, ..., n$, are i.i.d. m-dimensional random vectors following a multivariate normal distribution $\mathbf{N}(0, \mathbf{\Sigma})$ with the nonsingular covariance matrix $\mathbf{\Sigma}$.

Our goal is to estimate $\mathbf{B}$ so that we can use $\mathbf{X}$ to predict $\mathbf{Y}$. A simple way to estimate $\mathbf{B}$ is to build m single response models separately and the least squares solution is denoted by $\hat{\mathbf{B}}_S = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, provided that $\mathbf{X}^T\mathbf{X}$ is nonsingular. However, this approach ignores information on $\mathbf{\Sigma}$. When $\mathbf{\Sigma}$ is diagonal, this separate modeling approach can work well. However, when $\mathbf{\Sigma}$ is not diagonal, we sometimes have strong correlations among the response variables. The separate modeling approach does not make use of the joint information among the response variables. To produce a better estimator, we consider to incorporate $\mathbf{\Sigma}$ in the estimation procedure of $\mathbf{B}$. Denote $\mathbf{\Sigma}^{-1}$ by $\mathbf{\Omega}$. If we assume that $\mathbf{\Sigma}$ is known, the log-likelihood for $\mathbf{B}$ conditional on $\mathbf{X}$ is

$$-\frac{1}{2}\text{tr}\left\{(\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}(\mathbf{Y} - \mathbf{XB})^T\right\}, \tag{2.1}$$

up to a constant not depending on $\mathbf{B}$. Interestingly, although the maximum likelihood function involves $\mathbf{\Sigma}$, the corresponding maximum likelihood estimate turns out to be identical to the least squares estimate using the separate maximum likelihood method. This implies

that the maximizer of (2.1) does not take any advantage from the known information on $\mathbf{\Sigma}$. However, when we impose penalties on the likelihood, the joint method can bring some advantage in estimation. In this section, we propose to build multivariate regression models through joint shrinkage. The goal is to utilize the joint information among the m response variables to improve estimation and prediction. Since $\mathbf{\Sigma}$ is involved in the joint estimation and it is often unknown, we consider three different approaches: two plug-in approaches and the doubly penalized approach. The plug-in approach in Section 2.2.1 uses some estimator $\hat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$ to plug in the penalized likelihood function and then estimate $\mathbf{B}$ jointly. The plug-in approach in Section 2.2.2 estimates $\mathbf{\Omega}$ after plugging in a reasonable estimator of $\mathbf{B}$. The doubly penalized approach in Section 2.2.3 estimates $\mathbf{\Omega}$ and $\mathbf{B}$ simultaneously via regularizing the estimation of both $\mathbf{\Omega}$ and $\mathbf{B}$.

For discussion, we first assume that $\mathbf{\Sigma}$ is known. To regress $\mathbf{Y}$ on $\mathbf{X}$, we can model them separately, such as applying the LASSO for m different responses. Alternatively, we can use joint shrinkage estimation for the m response variables simultaneously. To demonstrate the difference between separate shrinkage and joint shrinkage, we consider a simple toy example for illustration. Suppose that $m = 2$, $p = 1$, and $\mathbf{X}^T\mathbf{X} = 1$. Let $\hat{\mathbf{B}}_S = (\hat{\beta}^S_{11}, \hat{\beta}^S_{12})$ be the least squares solution and assume that both $\hat{\beta}^S_{11}$ and $\hat{\beta}^S_{12}$ are positive and $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. With the penalty parameter λ, the separate LASSO solution is given by

$$\begin{aligned} \hat{\beta}^{LASSO}_{1m} &= \underset{\beta_{1m}}{\operatorname{argmin}} \left\{ (\mathbf{y}^m - \mathbf{X}\beta_{1m})^T (\mathbf{y}^m - \mathbf{X}\beta_{1m}) + \lambda |\beta_{1m}| \right\} \\ &= [\hat{\beta}^S_{1m} - \frac{\lambda}{2}]_+; \quad m = 1, 2, \end{aligned} \tag{2.2}$$

where $[u]_+ = u$ if $u \geq 0$ and $[u]_+ = 0$ if $u < 0$. In the joint shrinkage estimation, however, the solution is given by

$$\underset{\mathbf{B}}{\operatorname{argmin}} \left[\operatorname{tr} \left\{ (\mathbf{Y} - \mathbf{XB}) \mathbf{\Omega} (\mathbf{Y} - \mathbf{XB})^T \right\} + \lambda |\beta_{11}| + \lambda |\beta_{12}| \right]. \tag{2.3}$$

We can show that (2.3) is equivalent to

$$\underset{\mathbf{B}}{\operatorname{argmin}} \left[(\mathbf{B} - \hat{\mathbf{B}}_S) \mathbf{\Omega} (\mathbf{B} - \hat{\mathbf{B}}_S)^T + \lambda |\beta_{11}| + \lambda |\beta_{12}| \right] \tag{2.4}$$

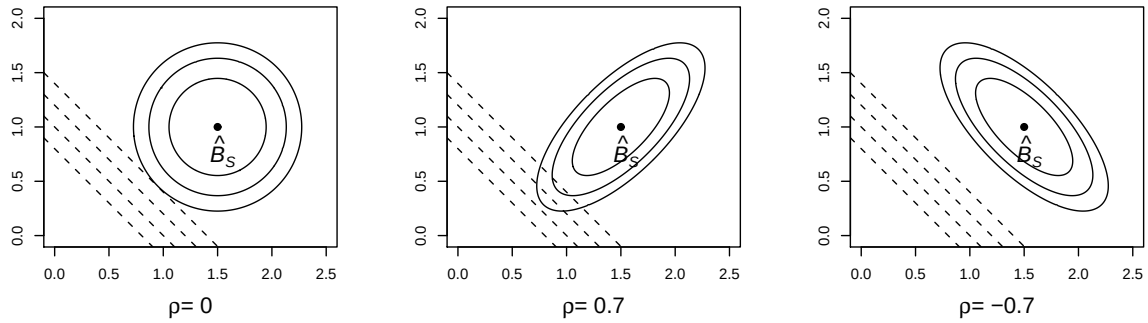

Figure 2.1: Contour plots for toy example to illustrate the change of shrinkage with ρ for the joint method.

and the solution of (2.4) is given by

$$\hat{\beta}_{1m} = [\hat{\beta}^S_{1m} - \frac{\lambda}{2}(1+\rho)]_+; \quad m = 1, 2. \tag{2.5}$$

Compared with the separate LASSO solution (2.2), the solution (2.5) obtains more shrinkage if ρ is positive, while negative ρ results in less shrinkage. Figure 2.1 provides some insight on the reason why the amount of shrinkage changes with ρ for the joint method. Solid curves in Figure 2.1 are contour curves of $(\mathbf{B} - \hat{\mathbf{B}}_S)\mathbf{\Omega}(\mathbf{B} - \hat{\mathbf{B}}_S)^T$ as the quadratic function of $\mathbf{B}$ and dashed lines correspond to the penalty function. When ρ is positive, the quadratic function increases along the 45° line to the horizontal axis slower than the case when ρ is zero. Note that the solution of the joint method with $\rho = 0$ is identical to the separate LASSO solution. Thus, the solution of (2.4) can be closer to the origin with more shrinkage than the solution with $\rho = 0$. On the other hand, the quadratic function with negative ρ increases faster along the 45° line to the horizontal axis. Thus, the solution of (2.4) tends to be closer to the least squares solution than the solution with $\rho = 0$. Therefore, the joint method can help us to produce more accurate estimators via utilizing the joint information through $\mathbf{\Omega}$.

We propose three approaches, including two plug-in methods and one joint method. In Sections 2.2.1 and 2.2.2, 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 $\mathbf{\Omega}$ prior to the step of

regression and then use the estimator of $\mathbf{\Omega}$ to produce a better estimator of $\mathbf{B}$. In the plug-in method for inverse covariance estimation, we estimate $\mathbf{B}$ first and then estimate $\mathbf{\Omega}$ with the estimator $\hat{\mathbf{B}}$ available. In Section 2.2.3, we estimate $\mathbf{B}$ and $\mathbf{\Omega}$ together via double penalization. Section 2.2.4 provides some guidance on three proposed methods and model selection.

2.2.1 Plug-in Joint Weighted LASSO Estimator

To ensure that estimation of $\mathbf{B}$ includes the information on $\mathbf{\Sigma}$, we propose a joint penalized likelihood method, namely the plug-in joint weighted LASSO (PWL) estimator. In particular, the corresponding penalized likelihood function is as follows

$$\operatorname{tr}\left\{(\mathbf{Y}-\mathbf{XB})\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB})^T\right\}+\lambda_1\sum_{j,k}w_{jk}\left|\beta_{jk}\right|. \tag{2.6}$$

Here λ_1 is a tuning parameter and $w_{jk}\geq 0$; $j=1,...,p$, $k=1,...,m$, are prespecified weights for the L_1-penalty of β_{jk}. If $\mathbf{\Omega}$ is an $m\times m$ diagonal matrix with diagonal entries $(\sigma_1^2,...,\sigma_m^2)$, then $\mathbf{y}^1,...,\mathbf{y}^m$ are mutually independent. In that case, the minimizer of (2.6) is equivalent to the weighted LASSO solution obtained by applying the weighted LASSO separately to each response vector $\mathbf{y}^k$ with the penalty parameter λ_1/σ_k^2 $(k=1,...,m)$. However, if $\mathbf{\Omega}$ is not diagonal, the minimizer of (2.6) can be different from the separate penalized likelihood method which handles each response vector $\mathbf{y}^k$ separately. Our numerical examples indicate that the joint method can be more accurate when the response variables are highly correlated.

In practice, $\mathbf{\Omega}$ is often not available. Thus, we need to estimate it. To estimate $\mathbf{\Omega}$, we assume that $\mathbf{z}_i=(\mathbf{y}_i^T,\mathbf{x}_i^T)^T$ is an $(m+p)$-dimensional random vector following a multivariate normal distribution $\mathbf{N}(\mu,\mathbf{\Sigma_{y,x}})$, where $\mathbf{\Sigma_{y,x}}=\begin{pmatrix}\Sigma_{y,y} & \Sigma_{y,x}\\ \Sigma_{x,y} & \Sigma_{x,x}\end{pmatrix}$. Because $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{y}_i$ conditioned on $\mathbf{x}_i$, it can be expressed by $\mathbf{\Sigma}=\Sigma_{y,y}-\Sigma_{y,x}\Sigma_{x,x}^{-1}\Sigma_{x,y}$. Therefore, we can estimate $\mathbf{\Sigma}$ by first estimating $\mathbf{\Sigma_{y,x}}$. To estimate $\mathbf{\Sigma_{y,x}}$, we adapt the Graphical LASSO (GLASSO) method proposed by Friedman, Hastie and Tibshirani [2008]. The GLASSO method considers the problem of estimating the inverse covariance matrix in the context of sparse Gaussian graphical models [Meinshausen and Buhlmann, 2006]. This technique

was also considered by Yuan and Lin [2007], Banerjee, Ghaoui and d'Aspremont [2008] and Rothman et al. [2008].

The GLASSO estimator, $\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{y},\mathbf{x}}$, is given as the minimizer of the following penalized likelihood function

$$-\log\det(\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}{}^{-1}) + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}{}^{-1}(\mathbf{z}_i - \bar{\mathbf{z}}) + \lambda_0\left\|\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}{}^{-1}\right\|. \tag{2.7}$$

Here $\bar{\mathbf{z}}$ is the sample mean, $\left\|\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}{}^{-1}\right\|$ is the sum of the absolute values of the off-diagonal elements of $\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}{}^{-1}$, and λ_0 is a tuning parameter.

The PWL method is a two-step procedure. With the estimate $\hat{\mathbf{\Sigma}}$ available, the PWL method solves the following problem

$$\underset{\mathbf{B}}{\operatorname{argmin}}\left[\operatorname{tr}\left\{(\mathbf{Y}-\mathbf{XB})\hat{\mathbf{\Omega}}(\mathbf{Y}-\mathbf{XB})^T\right\} + \lambda_1\sum_{j,k}w_{jk}\left|\beta_{jk}\right|\right], \tag{2.8}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{y},\mathbf{x}} = \begin{pmatrix} \hat{\Sigma}_{y,y} & \hat{\Sigma}_{y,x} \\ \hat{\Sigma}_{x,y} & \hat{\Sigma}_{x,x} \end{pmatrix}$, $\hat{\mathbf{\Sigma}} = \hat{\Sigma}_{y,y} - \hat{\Sigma}_{y,x}\hat{\Sigma}^{-1}_{x,x}\hat{\Sigma}_{x,y}$ and $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}}^{-1}$.

2.2.2 Plug-in Weighted Graphical LASSO Estimator

In Section 2.2.1, we propose a plug-in method, PWL, which estimates $\mathbf{\Omega}$ first and then estimates $\mathbf{B}$ given $\hat{\mathbf{\Omega}}$. In this section, we propose another plug-in method to estimate $\mathbf{\Omega}$. In particular, we first estimate $\mathbf{B}$ by using univariate regression techniques. With the estimator $\hat{\mathbf{B}}$ available, we propose a penalized likelihood method, the plug-in weighted graphical LASSO (PWGL) estimator, by solving

$$\underset{\mathbf{\Omega}}{\operatorname{argmin}}\left[-n\log\det(\mathbf{\Omega}) + \operatorname{tr}\left\{(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})\mathbf{\Omega}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T\right\} + \lambda_2\sum_{s\neq t}v_{st}|\omega_{st}|\right], \tag{2.9}$$

where $\mathbf{\Omega} = \{\omega_{st}\}; s = 1,...,m, t = 1,...,m$. Here λ_2 is a tuning parameter and $v_{st} \geq 0; s = 1,...,m, t = 1,...,m$, are prespecified weights for the L_1 penalty of ω_{st}.

2.2.3 Doubly Penalized Maximum Likelihood Estimator

In Sections 2.2.1 and 2.2.2, we propose two plug-in methods. PWL estimates $\mathbf{\Omega}$ first and then estimates $\mathbf{B}$ given $\hat{\mathbf{\Omega}}$ while PWGL estimates $\mathbf{B}$ first and then estimates $\mathbf{\Omega}$ given $\hat{\mathbf{B}}$. In

this section, we propose to estimate $(\mathbf{B}, \mathbf{\Omega})$ simultaneously. Since $\mathbf{y}_i|\mathbf{x}_i \sim \mathbf{N}(\mathbf{B}^T\mathbf{x}_i, \mathbf{\Sigma})$, the log-likelihood of $(\mathbf{B}, \mathbf{\Omega})$ conditional on $\mathbf{X}$ is

$$\frac{n}{2}\log\det(\mathbf{\Omega}) - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}-\mathbf{XB})\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB})^T\right\}. \tag{2.10}$$

It can be shown that the maximum likelihood estimator of $\mathbf{B}$ is also given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Interestingly, the resulting estimator of $\mathbf{B}$ is the same as the ordinary least square estimator, which can be obtained without using the information on the relationship among the response vectors $\mathbf{y}^1, ..., \mathbf{y}^m$. To incorporate the information among different response variables in estimation of $\mathbf{B}$, we propose a joint penalized method, the doubly penalized maximum likelihood (DML) estimator, by solving

$$\underset{\mathbf{B},\mathbf{\Omega}}{\mathrm{argmin}}\left[-n\log\det(\mathbf{\Omega}) + \mathrm{tr}\left\{(\mathbf{Y}-\mathbf{XB})\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB})^T\right\} + \lambda_1\sum_{j,k} w_{jk}\left|\beta_{jk}\right| + \lambda_2\sum_{s\neq t} v_{st}\left|\omega_{st}\right|\right]. \tag{2.11}$$

Note that the global minimizer of the objective function in (2.11) may not exist when $p \geq n$. This is because the first term in (2.11) can dominate the other terms if some diagonal elements of $\mathrm{tr}\left\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right\}$ are zeros, which may occur when $p \geq n$. This can be shown by taking a diagonal matrix $\mathbf{\Omega}$ and increasing the values of its diagonal elements corresponding to the zero diagonal entries in $\mathrm{tr}\left\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right\}$. As a result, the numerical solution of $\mathbf{\Omega}$ in (2.11) can have some large diagonal entries. In practice, the solution of $\mathbf{\Omega}$ with very large diagonal entries is not desirable as it implies the very small residual variances of the corresponding response variables. We recommend to first use the plug-in method in Section 2.2.1 or separate modeling methods to screen the variables and reduce the dimensions. Then one can apply the joint method on the reduced set of variables. As shown in our simulation examples, the joint method can often outperform the plug-in methods when p is moderate compared to n.

2.2.4 Model Selection

Two plug-in methods take advantages over the joint method if one of $\mathbf{B}$ and $\mathbf{\Omega}$ is of main interest and the other is already well estimated. Another advantage of two plug-in methods

is that they have less computational burden than the joint method. On the other hand, the joint method do not require good estimate of $\mathbf{B}$ or $\mathbf{\Omega}$. Even though the joint method is computationally more intensive, it often performs better than two plug-in methods in sense that it optimizes the log-likelihood of $(\mathbf{B}, \mathbf{\Omega})$ jointly.

The tuning parameters λ_1 and λ_2 in (2.8), (2.9) and (2.11) control the sparsity of the resulting estimators of $(\mathbf{B}, \mathbf{\Omega})$. They can be selected either using validation sets or through K-fold cross-validation. The K-fold cross-validation method randomly splits the dataset into K segments of equal sizes. For the k-th fold, we denote the estimated regression parameter matrix and the estimated inverse covariance matrix using all data excluding those in the k-th segment and the tuning parameters λ_1 and λ_2 by $(\hat{\mathbf{B}}_{\lambda_1}^{(-k)}, \hat{\mathbf{\Omega}}_{\lambda_2}^{(-k)})$. We also denote the data in the k-th segment as $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)})$. Specifically, for the PWL method, we select the optimal tuning parameter $\hat{\lambda}_1$ which minimizes the prediction error as follows,

$$\mathrm{CV}(\lambda_1) = \sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)} \hat{\mathbf{B}}_{\lambda_1}^{(-k)}||_F^2, \tag{2.12}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix. For the PWGL method, we select the optimal tuning parameter $\hat{\lambda}_2$ which minimizes the predictive negative log-likelihood as follows,

$$\mathrm{CV}(\lambda_2) = \sum_{k=1}^{K} \left[-n_k \log\det(\hat{\mathbf{\Omega}}_{\lambda_2}^{(-k)}) + \mathrm{tr}\left\{ (\mathbf{Y}^{(k)} - \mathbf{X}^{(k)} \hat{\mathbf{B}}) \hat{\mathbf{\Omega}}_{\lambda_2}^{(-k)} (\mathbf{Y}^{(k)} - \mathbf{X}^{(k)} \hat{\mathbf{B}})^T \right\} \right], \tag{2.13}$$

where n_k is the sample size of the k-th segment. For the DML method, we first select the optimal $\hat{\lambda}_1$ by using (2.12) with a prespecified λ_2 and select $\hat{\lambda}_2$ by using (2.13) with the selected optimal $\hat{\lambda}_1$. It helps to avoid a two dimensional grid search of (λ_1, λ_2). We have found in simulations that the selected optimal $\hat{\lambda}_1$s are almost identical for a wide range of prespecified λ_2.

In the use of validation sets, we split the dataset into two part, the training set and the validation set. With a pair of (λ_1, λ_2), we first estimate $(\mathbf{B}, \mathbf{\Omega})$ using the training set. The prediction error and the predictive negative log-likelihood of the resulting estimator are obtained using the validation set as $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)})$ in (2.12) and (2.13). The validation set is not used to construct the final estimator with the selected $(\hat{\lambda}_1, \hat{\lambda}_2)$ while the K-fold cross-validation uses all data for the final estimator with $(\hat{\lambda}_1, \hat{\lambda}_2)$.

2.3 Asymptotic Properties

To investigate a sparse regression technique, it is necessary to investigate its asymptotic behaviors. Fan and Li [2001] pointed out that a good variable selection procedure should have oracle properties. Asymptotically with probability tending to 1, a procedure with oracle properties can identify the true underlying subset of predictor variables. The resulting estimator of the procedure also asymptotically performs as well as if the true underlying subset were known in advance. In this section, we study the asymptotic behavior of our three proposed methods. In particular, we show that with a proper choice of (λ_1, λ_2), all three methods enjoy the oracle properties.

For the asymptotic analysis, we use the set-up of Fan and Li [2001], Yuan and Lin [2007] and Zou [2006]. The technical derivation uses the results in Knight and Fu [2000]. Let $\mathbf{B}^* = (\beta^*_{jk}); j = 1, ..., p, k = 1, ..., m$, be the true regression parameter matrix and $\mathbf{\Omega}^* = (\omega^*_{st}); s = 1, ..., m, t = 1, ..., m$, be the true inverse covariance matrix. Let $\mathcal{A} = \left\{(j,k) : \beta^*_{jk} \neq 0\right\}$ and $\mathcal{C} = \{(s,t) : \omega^*_{st} \neq 0\}$. Then we assume the following conditions for our theoretical results:

(A1) $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to A$ where A is a positive definite matrix.

(A2) The cardinality of $\mathcal{A}$, $|\mathcal{A}| = q_1 > 0$.

(A3) There exists $\tilde{\beta}_{jk}$ which is a $\sqrt{n}$-consistent estimator of $\beta^*_{jk}; j = 1, ..., p, k = 1, ..., m$.

(A4) The cardinality of $\mathcal{C}$, $|\mathcal{C}| = q_2 > 0$.

(A5) There exists $\tilde{\omega}_{st}$ which is a $\sqrt{n}$-consistent estimator of $\omega^*_{st}; s = 1, ..., m, t = 1, ..., m$.

Note that conditions **(A3)** and **(A5)** are generally satisfied by maximum likelihood estimators or L_2 regularized maximum likelihood estimators with proper choices of penalty parameters. For example, the least square estimator of $\mathbf{B}$ can be used as the $\tilde{\beta}_{jk}$s and the inverse of residual sample covariance matrix can be used as $\tilde{\omega}_{st}$s. For the theoretical analysis, we define w_{jk} and v_{st} as $w_{jk} = \frac{1}{|\tilde{\beta}_{jk}|^\gamma}; j = 1, ..., p, k = 1, ..., m$ where $\gamma > 0$ and $v_{st} = \frac{1}{|\tilde{\omega}_{st}|}; s = 1, ..., m, t = 1, ..., m$ respectively.

In Sections 2.3.1 and 2.3.2, we show the plug-in estimators enjoy the oracle properties. Section 2.3.3 develops the asymptotic theory that reveals the oracle properties of the DML solution.

2.3.1 Oracle properties of the PWL solution

In this section, we first show that with the known $\mathbf{\Omega}^*$, the minimizer of (2.6) is consistent in variable selection and has the asymptotic normality. Then we show that with a consistent estimator of $\mathbf{\Omega}^*$, the PWL estimator also enjoys the same properties.

Define the true regression parameter vector as $\boldsymbol{\beta}^* = (\beta^*_{11}, ..., \beta^*_{p1}, ..., \beta^*_{1m}, ..., \beta^*_{pm})^T$. Let $\hat{\boldsymbol{\beta}}^{\mathbf{1}(n)}$ be the estimator of $\boldsymbol{\beta}^*$ obtained by minimizing (2.6) with the penalty parameter $\lambda_{1,n}$. Let $\boldsymbol{\beta}^*_{\mathcal{A}}$ be the q_1-dimensional true parameter vector which consists of nonzero components in $\boldsymbol{\beta}^*$. Let $\hat{\boldsymbol{\beta}}^{\mathbf{1}(n)}_{\mathcal{A}}$ be the corresponding estimators of $\boldsymbol{\beta}^*_{\mathcal{A}}$. Let $D = (\mathbf{\Omega}^* \otimes A)_{\mathcal{A}}$ be the $q \times q$ matrix obtained by removing the $(j + (k-1)m)$-th row and column of $\mathbf{\Omega}^* \otimes A$ for $(j,k) \notin \mathcal{A}$. Then the following lemma shows the oracle properties of the penalized likelihood estimator $\hat{\boldsymbol{\beta}}^{\mathbf{1}(n)}$ with the known $\mathbf{\Omega}^*$, as the minimizer of (2.6) defined previously.

lemma 1. *(Oracle properties of the minimizer of* (2.6)*,* $\hat{\boldsymbol{\beta}}^{\mathbf{1}(n)}$*, with the known* $\mathbf{\Omega}^*$*) Suppose that* $\lambda_{1,n} n^{-\frac{1}{2}} \to 0$ *and* $\lambda_{1,n} n^{\frac{\gamma-1}{2}} \to \infty$ *as* $n \to \infty$*. Under the conditions* **(A1)-(A3)***, we have the following results:*

1. *(Selection consistency)* $\lim_n P(\hat{\beta}^{1(n)}_{jk} = 0) = 1$ *if* $\beta^*_{jk} = 0$ *;*
2. *(Asymptotic normality)* $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathbf{1}(n)}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d \mathbf{N}(0, D^{-1})$.

Lemma 1 tells us that the penalized maximum likelihood estimator with the known $\mathbf{\Omega}^*$ satisfies the oracle properties. Since $\mathbf{\Omega}^*$ is typically unknown in practice, one often uses an estimator for $\mathbf{\Omega}^*$. With slight modification of Lemma 1, we can show that the PWL solution also enjoys the oracle properties. Denote the PWL estimator of $\boldsymbol{\beta}^*$ with the penalty parameter $\lambda_{1,n}$ as $\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)}$. Let $\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)}_{\mathcal{A}}$ be the corresponding estimator of $\boldsymbol{\beta}^*_{\mathcal{A}}$.

Theorem 1. *(Oracle properties of the PWL solution) In addition to the assumptions in Lemma 1, suppose that* $\hat{\mathbf{\Omega}}$ *is a consistent estimator of* $\mathbf{\Omega}^*$*. Under the conditions* **(A1)-(A3)***, we have the following results:*

1. *(Selection consistency)* $\lim_n P(\hat{\beta}^{2(n)}_{jk} = 0) = 1$ *if* $\beta^*_{jk} = 0$ *;*
2. *(Asymptotic normality)* $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d \mathbf{N}(0, D^{-1})$.

Theorem 1 states that with a consistent estimator of $\mathbf{\Omega}^*$, variable selection in the PWL is consistent and the resulting estimator still enjoys the asymptotic normality.

2.3.2 Oracle properties of the PWGL solution

In this section, we show the oracle properties of the PWGL solution. To this end, we first show the oracle properties of the solution of

$$\underset{\mathbf{\Omega}}{\operatorname{argmin}} \left[-n \log\det(\mathbf{\Omega}) + \operatorname{tr}\left\{(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)\mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T\right\} + \lambda_2 \sum_{j \neq k} v_{jk}|\omega_{jk}| \right], \tag{2.14}$$

with the known $\mathbf{B}^*$. Then we show that with a consistent estimator of $\mathbf{B}^*$, the PWGL estimator still enjoys the same properties.

Denote by $\hat{\mathbf{\Omega}}^{(1)}$ the minimizer of (2.14) with the known $\mathbf{B}^*$. Let $\hat{\mathbf{\Omega}}_0^{(1)}$ be the matrix obtained from $\hat{\mathbf{\Omega}}^{(1)}$ by replacing $\hat{\omega}_{jk}^{(1)}$ with 0 if $\omega_{jk}^* = 0$. Then the following lemma shows the oracle properties of $\hat{\mathbf{\Omega}}^{(1)}$.

lemma 2. *(Oracle properties of the minimizer of* (2.14)*,* $\hat{\mathbf{\Omega}}^{(1)}$*, with known* $\mathbf{B}^*$*) Suppose that* $\lambda_{2,n} n^{-\frac{1}{2}} \to 0$ *and* $\lambda_{2,n} \to \infty$ *as* $n \to \infty$*. Under the conditions* **(A1), (A4)** *and* **(A5)***, we have the following results:*

1. *(Selection consistency)* $\lim_n P(\hat{\omega}_{jk}^{(1)} = 0) = 1$ *if* $\omega_{jk}^* = 0$ *;*
2. *(Asymptotic normality)* $\sqrt{n}(\hat{\mathbf{\Omega}}_0^{(1)} - \mathbf{\Omega}^*) \to_d \arg\min V(U)$*,*
   where $V(U) = tr(U\mathbf{\Sigma}U\mathbf{\Sigma}) + tr(UW)$ *and* W *is an* $m \times m$ *random symmetric matrix such that* $vec(W) \sim \mathbf{N}(0, \Lambda)$ *in which* $\operatorname{cov}(w_{ij}, w_{kl}) = \operatorname{cov}(\epsilon_{1i}\epsilon_{1j}, \epsilon_{1k}\epsilon_{1l})$*. The minimum is taken over all symmetric matrices* U *satisfying* $u_{jk} = 0$ *if* $\omega_{jk}^* = 0$*.*

In Lemma 2, we show that the penalized maximum likelihood estimator with the known $\mathbf{B}^*$ satisfies the oracle properties. Since $\mathbf{B}^*$ is typically unknown in practice, one often applies an univariate regression technique to obtain an estimator for $\mathbf{B}^*$. With slight modification of Lemma 2, we can show that the PWGL solution also enjoys the oracle properties. Denote the PWGL estimator of $\mathbf{\Omega}^*$ with the penalty parameter $\lambda_{2,n}$ as $\hat{\mathbf{\Omega}}^{(2)}$. Let $\hat{\mathbf{\Omega}}_0^{(2)}$ be the matrix obtained from $\hat{\mathbf{\Omega}}^{(2)}$ by replacing $\hat{\omega}_{jk}^{(2)}$ with 0 if $\omega_{jk}^* = 0$. Then the following theorem shows the oracle properties of the PWGL estimator.

Theorem 2. *(Oracle properties of the PWGL solution) In addition to the assumptions in Lemma 2, suppose that* $\hat{\mathbf{B}}$ *is a consistent estimator of* $\mathbf{B}^*$*. Under the above conditions, we have the following results:*

1. *(Selection consistency)* $\lim_n P(\hat{\omega}_{jk}^{(2)} = 0) = 1$ *if* $\omega_{jk}^* = 0$ *;*

2. *(Asymptotic normality)* $\sqrt{n}(\hat{\mathbf{\Omega}}_0^{(2)} - \mathbf{\Omega}^*) \to_d \arg\min V(U)$,
where $V(U) = tr(U\mathbf{\Sigma}U\mathbf{\Sigma}) + tr(UW)$ *and* W *is an* $m \times m$ *random symmetric matrix such that* $vec(W) \sim \mathbf{N}(0, \Lambda)$ *in which* $\mathrm{cov}(w_{ij}, w_{kl}) = \mathrm{cov}(\epsilon_{1i}\epsilon_{1j}, \epsilon_{1k}\epsilon_{1l})$*. The minimum is taken over all symmetric matrices* U *satisfying* $u_{jk} = 0$ *if* $\omega_{jk}^* = 0$*.*

Theorem 2 states that with a consistent estimator of $\mathbf{B}^*$, the PWGL solution satisfies the oracle properties.

2.3.3 Oracle properties of the DML solution

In Sections 2.3.1 and 2.3.2, we establish the oracle properties of plug-in estimators. In this section, we explore oracle properties of the DML solution in which $(\hat{\mathbf{B}}, \hat{\mathbf{\Omega}})$ are obtained together. First, we show that with a proper choice of (λ_1, λ_2), there exists a $\sqrt{n}$-consistent local minimizer of (2.11). Then we show that this local minimizer enjoys the oracle properties as a solution of the DML estimator.

The following lemma shows the existence of a local minimizer of (2.11) which is $\sqrt{n}$-consistent.

lemma 3. *Suppose that* $\lambda_{1,n}n^{-\frac{1}{2}} \to 0$ *and* $\lambda_{2,n}n^{-\frac{1}{2}} \to 0$*. Under the conditions* **(A1)-(A5)***, there exists a local minimizer of* (2.11) *such that*

$$\left\|(vec(\hat{\mathbf{B}})^T, vec(\hat{\mathbf{\Omega}})^T)^T - (vec(\mathbf{B}^*)^T, vec(\mathbf{\Omega}^*)^T)^T\right\| = O_p(1/\sqrt{n}).$$

From Lemma 3, it is clear that there exists a $\sqrt{n}$-consistent doubly penalized maximum likelihood estimator. As the DML estimator of $(\mathbf{B}^*, \mathbf{\Omega}^*)$, denote by $(\hat{\mathbf{B}}^{(n)}, \hat{\mathbf{\Omega}})$ the $\sqrt{n}$-consistent local solution of (2.11) with the penalty parameter $(\lambda_{1,n}, \lambda_{2,n})$. Let $\hat{\beta}^{(n)} = \mathrm{vec}(\hat{\mathbf{B}}^{(n)})$ and let $\hat{\beta}_{\mathcal{A}}^{(n)}$ be the corresponding estimator of $\boldsymbol{\beta}_{\mathcal{A}}^*$. Let $\hat{\mathbf{\Omega}}_0$ be the matrix obtained from $\hat{\mathbf{\Omega}}$ by replacing $\hat{\omega}_{jk}$ with 0 if $\omega_{jk}^* = 0$. We now show that with a proper choice of (λ_1, λ_2), the DML estimator as this local minimizer enjoys the oracle properties in the following theorem.

Theorem 3. *(Oracle properties of the DML solution) Suppose that* $\lambda_{1,n}n^{-\frac{1}{2}} \to 0$ *and* $\lambda_{1,n}n^{\frac{\gamma-1}{2}} \to \infty$*. In addition to that, suppose that* $\lambda_{2,n}n^{-\frac{1}{2}} \to 0$ *and* $\lambda_{2,n} \to \infty$*. Under*

the conditions **(A1)-(A5)**, *we have the following results:*

1. $\lim_n P(\hat{\beta}_{jk}^{(n)} = 0) = 1$ *if* $\beta_{jk}^* = 0$;

2. $\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{(n)} - \boldsymbol{\beta}_{\mathcal{A}}^*) \to_d \mathbf{N}(0, D^{-1})$;

3. $\lim_n P(\hat{\omega}_{jk} = 0) = 1$ *if* $\omega_{jk}^* = 0$;

4. $\sqrt{n}(\hat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}^*) \to_d \arg\min V(U)$,
   where $V(U) = tr(U\boldsymbol{\Sigma}U\boldsymbol{\Sigma}) + tr(UW)$ *and* W *is a* $m \times m$ *random symmetric matrix such that* $vec(W) \sim \mathbf{N}(0, \Lambda)$ *in which* $\mathrm{cov}(w_{ij}, w_{kl}) = \mathrm{cov}(\epsilon_{1i}\epsilon_{1j}, \epsilon_{1k}\epsilon_{1l})$. *The minimum is taken over all symmetric matrices* U *satisfying* $u_{jk} = 0$ *if* $\omega_{jk}^* = 0$.

2.4 Computational Algorithm

In this section, we describe computational algorithms to solve problems (2.8), (2.9), and (2.11). In particular, we apply the GLASSO algorithm for (2.9). To solve the problems (2.8) and (2.11), we apply the coordinate-descent algorithm as described in Peng et al. [2009], which can be viewed as a modification of the shooting algorithm [Fu, 1998]. The basic idea of the coordinate-descent algorithm is to optimize each parameter at one time while holding the other parameters fixed at the current solution. The corresponding optimization at each step can be very simple to solve.

We now describe the coordinate-descent algorithm for the PWL method in details. Denote $\hat{\boldsymbol{\Omega}}$ by $(\hat{\omega}_{ij})_{m\times m}$. Then (2.8) is equivalent to minimizing

$$\sum_{i=1}^{n}\sum_{k,l=1}^{m}\hat{\omega}_{kl}(y_{ik} - \sum_{j=1}^{p}\beta_{jk}x_{ij})(y_{il} - \sum_{j=1}^{p}\beta_{jl}x_{ij}) + \lambda_1\sum_{j,k}w_{jk}\left|\beta_{jk}\right|. \tag{2.15}$$

Consider (2.15) as a function of β_{jk} with other coefficients fixed. Then the minimizer of (2.15) is equivalent to

$$\underset{\beta_{jk}}{\mathrm{argmin}}\left[\left\{\sum_{i=1}^{n}\left(\hat{\omega}_{kk}(y_{ik} - \sum_{j'\neq j}\beta_{j'k}x_{ij'} - \beta_{jk}x_{ij})^2 \right.\right.\right.$$
$$\left.\left.\left. +2\sum_{k'\neq k}\hat{\omega}_{kk'}(y_{ik'} - \sum_{j}\beta_{jk'}x_{ij})(y_{ik} - \sum_{j'\neq j}\beta_{j'k}x_{ij'} - \beta_{jk}x_{ij})\right)\right\} + \lambda_1 w_{jk}\left|\beta_{jk}\right|\right].$$

This problem is essentially a one-dimensional LASSO optimization which has a closed form solution. Therefore, the algorithm can be summarized as follows:

Algorithm 1: the Coordinate-Descent Algorithm for the PWL Method

Step 1 (Initial value). Set the separate LASSO solution $\beta_{jk}^{(old)}; j = 1, ..., p, k = 1, ..., m$, as the initial value for $\mathbf{B}$.

Step 2 (Updating rule). For $j = 1, ..., p$ and $k = 1, ..., m$,

$\beta_{qr}^{(new)} = \beta_{qr}^{(old)}$, if $q \neq j$ and $r \neq k$,

$$\beta_{jk}^{(new)} = \text{sign}\left(\frac{\sum_{l=1}^{m} \hat{\omega}_{lk}(e_l^{(old)})^T \mathbf{x^j}}{\hat{\omega}_{kk}\mathbf{x^j}^T\mathbf{x^j}} + \beta_{jk}^{(old)}\right)\left(\left|\frac{\sum_{l=1}^{m} \hat{\omega}_{lk}(e_l^{(old)})^T \mathbf{x^j}}{\hat{\omega}_{kk}\mathbf{x^j}^T\mathbf{x^j}} + \beta_{jk}^{(old)}\right| - \frac{\lambda_1 w_{jk}}{2\hat{\omega}_{kk}\mathbf{x^j}^T\mathbf{x^j}}\right)^+,$$

where $e_l^{(old)} = \mathbf{y}^l - \mathbf{X}\boldsymbol{\beta}^{l(old)}$ and $\boldsymbol{\beta}^{l(old)} = (\beta_{1l}^{(old)}, ..., \beta_{pl}^{(old)})$.

Step 3 (Iteration). Repeat Step 2 until convergence. Our stopping rule is that the change of the objective function in (2.8) is less than $\delta = 0.1$.

To be computationally more efficient, we combine the above algorithm with the active shooting algorithm proposed by Peng et al. [2009]. The basic idea of the active shooting algorithm is to update the coefficients within the active set until convergence instead of iterating all coefficients at each step. The active set is defined as the set of currently nonzero coefficients and it is typically small. Once the coefficients in the active set converge, then we continue to update other coefficients. This step can speed up the algorithm significantly if the final solution is very sparse.

Next we describe the problem (2.9) in the GLASSO framework. Since (2.9) is equivalent to minimizing

$$-\log\det(\boldsymbol{\Omega}) + \text{tr}\left\{\frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\boldsymbol{\Omega}\right\} + \frac{\lambda_2}{n}\sum_{j\neq k} v_{jk}|\omega_{jk}|, \tag{2.16}$$

we can apply the GLASSO algorithm [Friedman, Hastie and Tibshirani, 2008] to solve (2.9) by substituting the sample covariance matrix with $\frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$. Therefore, the algorithm for (2.9) proceeds as follows:

Algorithm 2: the GLASSO Algorithm for the PWGL Method

Step 1 (Estimator of $\mathbf{B}$) Set the separate LASSO solution as the estimator, $\hat{\mathbf{B}}$, of $\mathbf{B}$.

Step 2 (Estimator of $\boldsymbol{\Omega}$) Given $\hat{\mathbf{B}}$, apply the GLASSO algorithm to solve (2.16).

Next, we combine Algorithm 1 and the GLASSO algorithm to solve problem (2.11) for the doubly penalized method DML in Section 2.2.3. The algorithm can be summarized as follows:

Algorithm 3: the Coordinate-Descent Algorithm for the DML Method

Step 1 (Initial values of $\mathbf{B}$ and $\mathbf{\Omega}$). Set the separate LASSO solution $\beta_{jk}^{(old)}; j = 1, ..., p, k = 1, ..., m$, as the initial value for $\mathbf{B}$ and the solution of (2.9), $\mathbf{\Omega}^{(old)}$, as the initial value of $\mathbf{\Omega}$.

Step 2 ($\mathbf{B}$ updating rule). For a given $\mathbf{\Omega}^{(old)}$, update $\mathbf{B}^{(old)} \rightarrow \mathbf{B}^{(new)}$ with

$$\mathbf{B}^{(new)} = \underset{\mathbf{B}}{\operatorname{argmin}} \left[\operatorname{tr}\left\{ (\mathbf{Y} - \mathbf{XB})\mathbf{\Omega}^{(old)}(\mathbf{Y} - \mathbf{XB})^T \right\} + \lambda_1 \sum_{j,k} w_{jk} \left|\beta_{jk}\right| \right].$$

This step can be solved using the Algorithm 1.

Step 3 ($\mathbf{\Omega}$ updating rule). For a given $\mathbf{B}^{(new)}$, update $\mathbf{\Omega}^{(old)} \rightarrow \mathbf{\Omega}^{(new)}$ by

$$\mathbf{\Omega}^{(new)} = \underset{\mathbf{\Omega}}{\operatorname{argmin}} \left[\operatorname{tr}\left\{ \frac{1}{n}(\mathbf{Y} - \mathbf{XB}^{(new)})^T(\mathbf{Y} - \mathbf{XB}^{(new)})\mathbf{\Omega} \right\} - \log\det(\mathbf{\Omega}) + \frac{\lambda_2}{n} \sum_{s \neq t} v_{st} \left|\omega_{st}\right| \right].$$

This can be solved using the GLASSO algorithm.

Step 4 (Iteration). Repeat Steps 2 and 3 until convergence. Our stopping rule is that the change of the objective function in (2.11) is less than $\delta = 0.1$.

Based on our experiment, the coordinate-descent algorithm works very efficiently. Since the DML method involves estimation of both $\mathbf{B}$ and $\mathbf{\Omega}$, the computation can be intensive when the dimension is high. We consider a prescreening step to speed up the computation. In particular, we adapt the group lasso method considered by Yuan and Lin [2006] and Meier, van de Geer and Buhlmann [2008]. The basic idea of the group lasso method is to employ group penalty in the regression problem so that model selection can be achieved in terms of group selection. In our multiple response variable regression problem, $(\beta_{j1}, ..., \beta_{jm}); j = 1, ..., p$, can be considered as p groups. Therefore, for the prescreening step, the group lasso estimator, $\hat{\mathbf{B}}^{group}$ of $\mathbf{B}$, is given as the minimizer of the following penalized function

$$\sum_{i=1}^{n} \sum_{k=1}^{m} (y_{ik} - \sum_{j=1}^{p} \beta_{jk} x_{ij})^2 + \lambda \sum_{j=1}^{p} \sqrt{\beta_{j1}^2 + ... + \beta_{jm}^2},$$

where λ is a tuning parameter. We screen out a variable if the corresponding coefficients

are estimated as zeros for all response variables. In other words, we remove the variable $\mathbf{x}^j$ from our model if $\hat{\beta}_{j1}^{group} = ... = \hat{\beta}_{jm}^{group} = 0$. This prescreening step can not only improve the prediction performance as shown in our examples, but also speed up the computation.

2.5 Simulated Examples

In this section, we compare our proposed methods with several existing methods. The first existing method we compare is the curds and whey (CW) method proposed by Breiman and Friedman [1997]. We use the CW with the generalized cross validation (CW-GCV) when $p < n$ and the CW with the ridge regression (CW-RR) when $p \geq n$. The other two methods are the separate ridge regression (RR) and the separate LASSO. In particular, we apply the RR and the LASSO to each response variable separately. The LASSO solution is constructed by the LARS algorithm proposed by Efron et al. [2004].

For comparison, consider three simulated examples with different $\mathbf{\Omega}$ structures. In all examples, $\mathbf{y}_i$ is a 10-dimensional multivariate vector and the predictors $\mathbf{x}_i$ $(i = 1, ..., n)$ are i.i.d. normal vectors from $\mathbf{N}(0, I_p)$. For each example, we simulate samples with size n $(n = 40, 60, 100, 150, 200)$ and dimension p of predictors $(p = 20, 40)$. The true regression coefficients for predictors are the following:

- For $j = 13, ..., p$, $\beta_{j,k} = 0$ $(k = 1, .., 10)$, which makes predictor $\mathbf{x}^j$ random noise if $j \geq 13$.
- For $j = 1, ..., 10$, $\beta_{j,j} = 3$, $\beta_{j,j+1} = 4$, $\beta_{j,j+2} = 3$, and otherwise $\beta_{j,k} = 0.5$.

We consider the following three different $\mathbf{\Omega}$ structures:

- **Example 1**: Banded inverse covariance matrix

  $\omega_{i,i} = i/5$, $\omega_{i,i+1} = \omega_{i+1,i} = (i(i+1)/100)^{1/2}$ $(i = 1, .., 9)$, otherwise $\omega_{i,j} = 0$.

- **Example 2**: Sparse inverse covariance matrix

  $\omega_{i,i} = i/5$, $\omega_{i,j} = ((i \times j)/100)^{1/2}$ $(i \neq j,\ max(i,j) \leq 5)$, otherwise $\omega_{i,j} = 0$.

- **Example 3**: Non-sparse inverse covariance matrix

  $\omega_{i,i} = 1$, $\omega_{i,j} = 0.5$ $(i \neq j)$.

For illustration, we show inverse covariance structures of these three examples in Figure 2.2. The structure of the inverse covariance in Example 1 is banded. Example 2 has a nonzero block on the lower left corner and Example 3 has no zero entries in $\mathbf{\Omega}$.

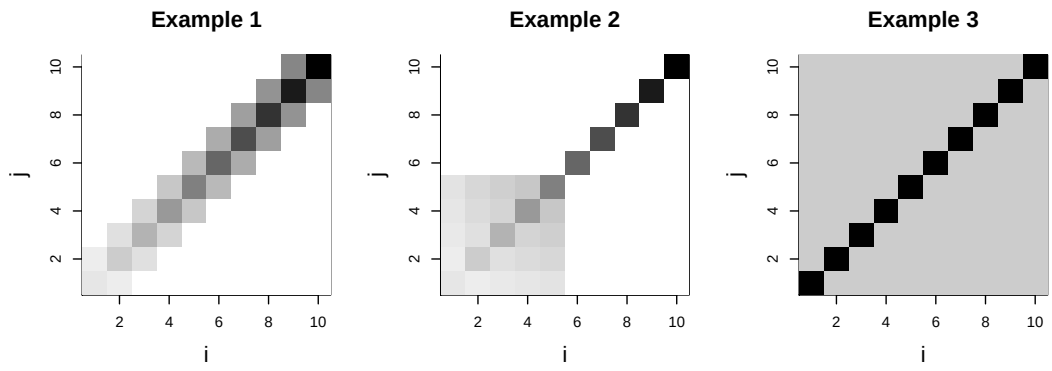

Figure 2.2: Inverse Covariance Structures for Examples 1-3. Example 1 has banded inverse covariance structures. The off-diagonal elements in Example 2 are zeros except the lower left block. All elements in Example 3 are non-zeros.

For the tuning parameter selection, we generate a tuning set with the same size n of the training set. There are two tuning parameters: λ_1 is for $\mathbf{B}$ estimation and λ_2 is for $\mathbf{\Omega}$ estimation. As the criterion of parameter selection for $\mathbf{B}$ estimation, we use the predictive squared error (PSE)

$$\text{PSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \text{tr}\{(\hat{\mathbf{Y}} - \mathbf{Y})^T(\hat{\mathbf{Y}} - \mathbf{Y})\}.$$

The PSE measures predictive accuracy with squared error. For the tuning parameter selection in $\mathbf{\Omega}$ estimation, we use the entropy criterion (Ent)[Huang et al., 2006] as follows:

$$Ent(\mathbf{\Omega}, \hat{\mathbf{\Omega}}) = \text{tr}(\mathbf{\Omega}^{-1}\hat{\mathbf{\Omega}}) - \log(|\mathbf{\Omega}^{-1}\hat{\mathbf{\Omega}}|) - m.$$

The Ent measures the difference of two matrices. For the RR estimator and the LASSO estimator, we apply the RR and the LASSO to each response vector $\mathbf{y}^k$ $(k = 1, ..., m)$ separately and select the tuning parameter which minimizes PSE in the tuning set. As mentioned in Section 2.4, the separate LASSO estimator is used as $\hat{\mathbf{B}}$ in our PWGL method. For our PWGL estimator, the tuning parameter is selected by minimizing Ent in the tuning set. For our PWL estimator, we first estimate $\mathbf{\Omega}$ by using the GLASSO. The tuning parameter of the GLASSO is selected by minimizing Ent in the tuning set. For the tuning parameter in $\mathbf{B}$ estimation, λ_1 is selected by minimizing PSE in the tuning set. Based on our observation, a reasonable choice of λ_2 is around [0.25, 1]. To avoid two-dimensional

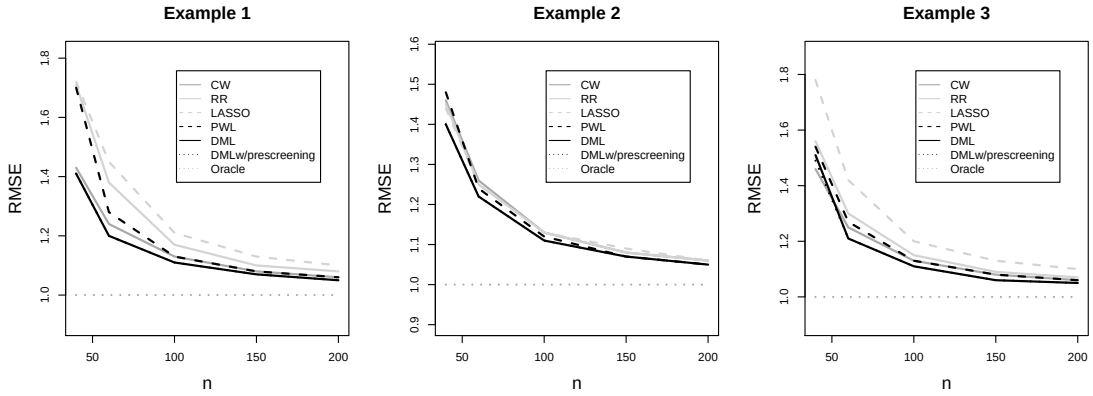

Figure 2.3: Averages of RMSE with p=20 for simulated Examples 1-3.

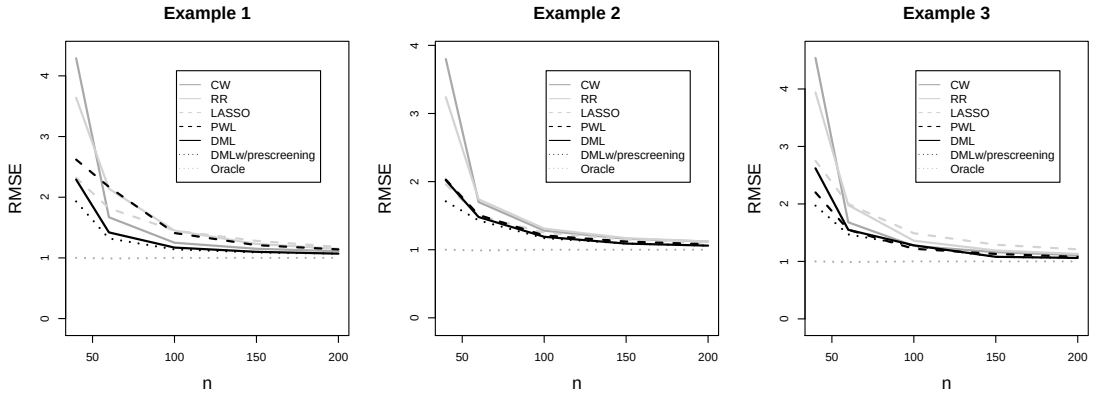

Figure 2.4: Averages of RMSE with p=40 for simulated Examples 1-3.

grid search of our DML method, we first select λ_1 which minimizes PSE in the tuning set while we fix λ_2 = 0.5. Then we select λ_2 which minimizes Ent in the tuning set with the selected λ_1. In all examples, we use $1/\sqrt{\beta_{jk}^{RR}}$ as the weights w_{jk} in our two proposed methods, PWL and DML, where β_{jk}^{RR} is the ridge regression estimator obtained by applying the ridge regression to each response variable $\mathbf{y}^k$ [Zou, 2006].

To compare prediction accuracy among methods, we report the standardized version of root mean square error (RMSE)

$$\left[\frac{1}{nm}\mathrm{tr}\{(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})\mathbf{\Omega}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T\}\right]^{1/2}.$$

We also report ratio of correctly identified zero coefficients among true zero coefficients in the $\mathbf{B}$ estimation step to compare the percentage of sparsity obtained. For the comparison of the $\mathbf{\Omega}$ estimation, we report the entropy, $Ent(\mathbf{\Omega}, \hat{\mathbf{\Omega}})$.

Table 2.1: Averages of RMSE and standard errors based on 100 replications (The numbers in parentheses are standard errors).

		n=40	n=60	n=100	n=150	n=200
p = 20	Example 1					
	CW	1.43 (0.009)	1.24 (0.004)	1.13 (0.003)	1.08 (0.002)	1.06 (0.002)
	RR	1.71 (0.023)	1.38 (0.010)	1.17 (0.005)	1.10 (0.003)	1.08 (0.002)
	LASSO	1.72 (0.024)	1.45 (0.014)	1.21 (0.007)	1.13 (0.005)	1.10 (0.003)
	PWL	1.70 (0.034)	1.28 (0.009)	1.13 (0.003)	1.08 (0.002)	1.06 (0.002)
	DML	1.41 (0.020)	1.20 (0.006)	1.11 (0.003)	1.07 (0.002)	1.05 (0.002)
	DMLw/prescreening	1.41 (0.021)	1.20 (0.006)	1.11 (0.003)	1.07 (0.002)	1.05 (0.002)
	Oracle	1.00 (0.003)	1.00 (0.002)	1.00 (0.003)	1.00 (0.002)	1.00 (0.002)
	Example 2					
	CW	1.46 (0.009)	1.26 (0.005)	1.13 (0.003)	1.08 (0.002)	1.06 (0.002)
	RR	1.45 (0.010)	1.25 (0.004)	1.13 (0.004)	1.08 (0.002)	1.06 (0.002)
	LASSO	1.44 (0.010)	1.25 (0.005)	1.13 (0.003)	1.09 (0.002)	1.06 (0.002)
	PWL	1.48 (0.016)	1.24 (0.005)	1.12 (0.003)	1.07 (0.002)	1.05 (0.002)
	DML	1.40 (0.009)	1.22 (0.005)	1.11 (0.003)	1.07 (0.002)	1.05 (0.002)
	DMLw/prescreening	1.40 (0.009)	1.22 (0.005)	1.11 (0.003)	1.07 (0.002)	1.05 (0.002)
	Oracle	1.00 (0.003)	1.00 (0.002)	1.00 (0.003)	1.00 (0.002)	1.00 (0.002)
	Example 3					
	CW	1.46 (0.009)	1.25 (0.005)	1.13 (0.003)	1.08 (0.002)	1.06 (0.002)
	RR	1.56 (0.015)	1.30 (0.006)	1.15 (0.004)	1.09 (0.002)	1.07 (0.002)
	LASSO	1.78 (0.022)	1.42 (0.009)	1.20 (0.005)	1.13 (0.003)	1.10 (0.003)
	PWL	1.54 (0.020)	1.27 (0.006)	1.13 (0.004)	1.08 (0.003)	1.06 (0.002)
	DML	1.51 (0.018)	1.21 (0.004)	1.11 (0.003)	1.06 (0.002)	1.05 (0.002)
	DMLw/prescreening	1.49 (0.017)	1.21 (0.004)	1.11 (0.003)	1.06 (0.002)	1.05 (0.002)
	Oracle	1.00 (0.003)	1.00 (0.002)	1.00 (0.003)	1.00 (0.002)	1.00 (0.002)
p = 40	Example 1					
	CW	4.29 (0.059)	1.67 (0.011)	1.25 (0.004)	1.15 (0.003)	1.11 (0.002)
	RR	3.64 (0.048)	2.14 (0.025)	1.45 (0.010)	1.23 (0.005)	1.15 (0.003)
	LASSO	2.32 (0.035)	1.82 (0.022)	1.45 (0.011)	1.28 (0.007)	1.18 (0.004)
	PWL	2.62 (0.045)	2.17 (0.019)	1.41 (0.011)	1.21 (0.004)	1.14 (0.003)
	DML	2.29 (0.042)	1.42 (0.014)	1.17 (0.003)	1.10 (0.002)	1.07 (0.002)
	DMLw/prescreening	1.93 (0.037)	1.32 (0.011)	1.14 (0.004)	1.09 (0.002)	1.07 (0.002)
	Oracle	1.00 (0.003)	0.99 (0.002)	1.00 (0.002)	1.00 (0.002)	1.00 (0.001)
	Example 2					
	CW	3.80 (0.069)	1.70 (0.011)	1.28 (0.004)	1.16 (0.003)	1.12 (0.002)
	RR	3.24 (0.062)	1.74 (0.014)	1.31 (0.004)	1.17 (0.003)	1.12 (0.002)
	LASSO	1.96 (0.032)	1.50 (0.013)	1.25 (0.005)	1.15 (0.003)	1.10 (0.002)
	PWL	2.03 (0.037)	1.51 (0.017)	1.21 (0.005)	1.12 (0.003)	1.08 (0.002)
	DML	2.02 (0.034)	1.48 (0.011)	1.19 (0.003)	1.09 (0.002)	1.06 (0.002)
	DMLEw/prescreening	1.71 (0.058)	1.43 (0.009)	1.17 (0.004)	1.09 (0.002)	1.06 (0.002)
	Oracle	1.00 (0.004)	0.99 (0.003)	1.00 (0.002)	1.00 (0.002)	1.00 (0.001)
	Example 3					
	CW	4.54 (0.091)	1.68 (0.011)	1.27 (0.004)	1.16 (0.003)	1.11 (0.002)
	RR	3.94 (0.082)	1.98 (0.023)	1.36 (0.006)	1.19 (0.003)	1.13 (0.002)
	LASSO	2.75 (0.050)	2.00 (0.027)	1.49 (0.009)	1.29 (0.006)	1.21 (0.004)
	PWL	2.20 (0.054)	1.55 (0.017)	1.22 (0.005)	1.13 (0.003)	1.08 (0.002)
	DML	2.62 (0.081)	1.55 (0.018)	1.28 (0.012)	1.08 (0.002)	1.06 (0.002)
	DMLw/prescreening	1.97 (0.080)	1.47 (0.015)	1.27 (0.013)	1.08 (0.002)	1.06 (0.002)
	Oracle	1.00 (0.003)	0.99 (0.003)	1.00 (0.002)	1.00 (0.002)	1.00 (0.001)

Table 2.1 reports the RMSE results for different settings and Figures 2.3-2.4 summarize the results. In terms of the RMSE criterion, our proposed method DML shows the best results in all examples. As n increases, performance of all methods gets closer to oracle, $\left[\frac{1}{nm}\text{tr}\{(\mathbf{Y}-\mathbf{XB})\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB})^T\}\right]^{1/2}$, which is the true underlying error. In Example 1 with $p = 20$, our proposed method PWL shows better performance than RR or LASSO. Although PWL is worse than CW when $n < 100$, it shows similar performance if $n \geq 100$. As p increases, PWL performs worse than LASSO when n is small. It indicates that GLASSO estimation of inverse covariance matrix may not be good enough when the sample size is very small. However, as n increase, it performs more competitively. Notice that DML shows best performance even when $p = 40$ and n is small. In Example 2 when $p = 20$, all methods give similar RMSEs. This is natural as the inverse covariance matrix in Example 2 is close to the diagonal matrix. Therefore, the separate approach and joint approach give similar results. When $p = 40$ and $n \leq 60$, RR and CW perform poorly although LASSO, PWL, and DML show similar performance as in the case when $p = 20$. This is because RR and CW estimators may not work well in ill-conditioned cases. In Example 3, the inverse covariance matrix is not sparse. LASSO gives the worst RMSE while the other methods show similar performance except the case when $n = p$. This implies that joint approaches outperform separate approaches in this case. Overall, the proposed DML method works the best in terms of the RMSE. The PWL method also works reasonably well in all cases, although it is not as accurate as the DML estimator. With respect to the prescreening step in the algorithm for the DML method, we can see that it even improves the prediction performance when $p = 40$ and n is small.

Table 2.2 reports ratios of correctly identified zero coefficients and they are summarized in Figures 2.5-2.6. When $p = 20$, Example 1 shows that DML outperforms LASSO and PWL in terms of ratios of correctly identified zero coefficients. As n increases, performances of three methods get closer. In Examples 2 and 3, DML and PWL identify zero coefficients more accuately than LASSO. We also notice that the ratios of correctly identified zeros for PWL and DML increase as n increases while the ratio for LASSO does not. The results support the selection consistency we have proved in Section 2.3. When $p = 40$, Examples 2-3 also show the better performance of PWL. In Examples 2-3, we notice that the ratios for DML tend to increase as n increases as we expected.

Table 2.2: Averages of ratio of correctly identified zero coefficients and standard errors based on 100 replications (The numbers in parentheses are standard errors).

		n=40	n=60	n=100	n=150	n=200
	Example 1					
	LASSO	0.45 (0.013)	0.39 (0.012)	0.30 (0.009)	0.27 (0.007)	0.27 (0.007)
	PWL	0.39 (0.019)	0.25 (0.013)	0.19 (0.010)	0.22 (0.011)	0.22 (0.010)
	DML	0.44 (0.013)	0.53 (0.010)	0.39 (0.014)	0.36 (0.016)	0.33 (0.014)
	DMLw/prescreening	0.49 (0.016)	0.56 (0.011)	0.41 (0.015)	0.38 (0.017)	0.36 (0.016)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)
	Example 2					
	LASSO	0.32 (0.010)	0.29 (0.009)	0.26 (0.007)	0.25 (0.007)	0.27 (0.007)
	PWL	0.61 (0.015)	0.57 (0.013)	0.59 (0.009)	0.65 (0.010)	0.69 (0.008)
p = 20	DML	0.40 (0.011)	0.45 (0.012)	0.46 (0.012)	0.48 (0.012)	0.53 (0.013)
	DMLw/prescreening	0.41 (0.012)	0.46 (0.012)	0.47 (0.012)	0.49 (0.012)	0.53 (0.013)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)
	Example 3					
	LASSO	0.36 (0.010)	0.29 (0.008)	0.25 (0.007)	0.25 (0.006)	0.27 (0.006)
	PWL	0.64 (0.014)	0.58 (0.012)	0.59 (0.010)	0.65 (0.008)	0.71 (0.011)
	DML	0.42 (0.010)	0.44 (0.009)	0.46 (0.008)	0.52 (0.007)	0.59 (0.010)
	DMLw/prescreening	0.43 (0.011)	0.44 (0.009)	0.46 (0.008)	0.53 (0.008)	0.59 (0.010)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)
	Example 1					
	LASSO	0.67 (0.007)	0.61 (0.009)	0.54 (0.010)	0.51 (0.008)	0.49 (0.008)
	PWL	0.79 (0.006)	0.64 (0.006)	0.42 (0.010)	0.38 (0.007)	0.38 (0.007)
	DML	0.49 (0.012)	0.53 (0.011)	0.52 (0.012)	0.56 (0.012)	0.55 (0.011)
	DMLw/prescreening	0.72 (0.012)	0.64 (0.010)	0.66 (0.013)	0.68 (0.012)	0.62 (0.010)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)
	Example 2					
	LASSO	0.59 (0.007)	0.53 (0.008)	0.49 (0.008)	0.48 (0.006)	0.49 (0.008)
	PWL	0.87 (0.006)	0.84 (0.006)	0.81 (0.005)	0.81 (0.006)	0.85 (0.004)
p = 40	DML	0.57 (0.009)	0.40 (0.006)	0.56 (0.012)	0.69 (0.004)	0.68 (0.003)
	DMLw/prescreening	0.51 (0.019)	0.47 (0.010)	0.67 (0.011)	0.70 (0.006)	0.69 (0.005)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)
	Example 3					
	LASSO	0.62 (0.006)	0.55 (0.008)	0.49 (0.008)	0.47 (0.007)	0.48 (0.007)
	PWL	0.87 (0.006)	0.83 (0.007)	0.78 (0.007)	0.85 (0.006)	0.82 (0.003)
	DML	0.56 (0.013)	0.41 (0.008)	0.59 (0.013)	0.69 (0.007)	0.69 (0.003)
	DMLw/prescreening	0.53 (0.022)	0.48 (0.012)	0.64 (0.011)	0.70 (0.006)	0.71 (0.005)
	Oracle	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)	1.00 (0.000)

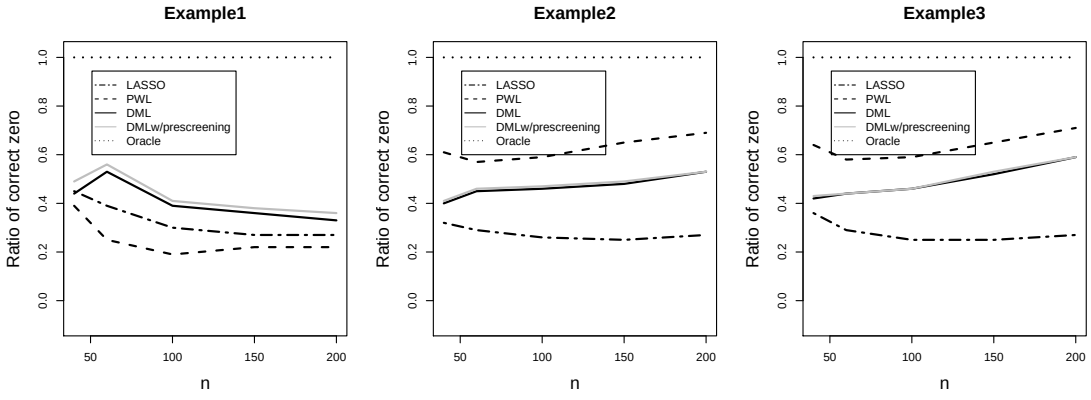

Figure 2.5: Averages of ratio of correctly identified zero coefficients with p=20.

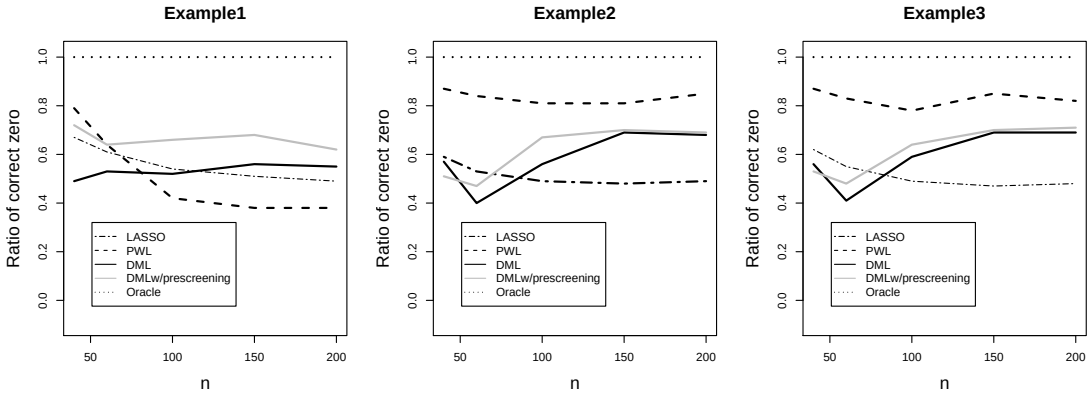

Figure 2.6: Averages of ratio of correctly identified zero coefficients with p=40.

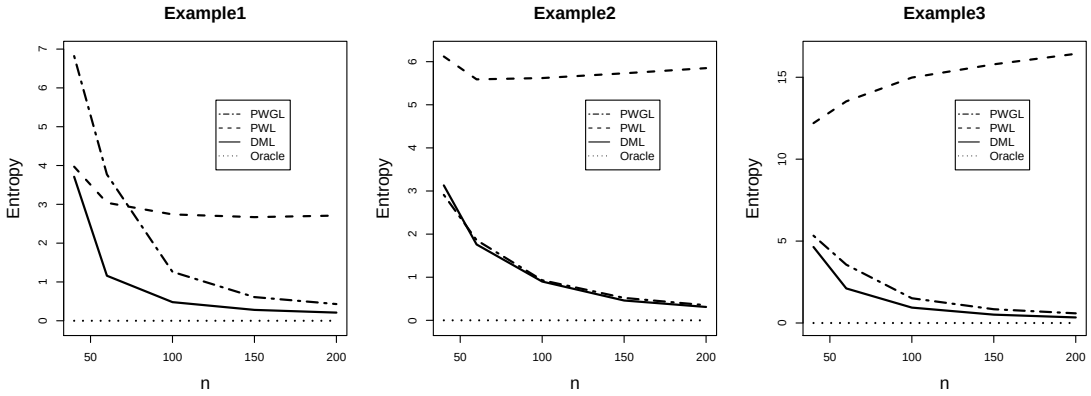

Figure 2.7: Averages of Entropy with p=20.

Figures 2.7-2.8 summarize the entropy results of two methods: PWGL and DML. In all examples when p = 20, the DML estimator shows the best performance. When p = 40 and $n \leq 60$, PWGL outperforms DML. Since the DML method simultaneously estimates both $\mathbf{B}$ and $\mathbf{\Omega}$, with a small n, the $\mathbf{\Omega}$ estimation may not be as good. However, as n increases, DML outperforms PWGL in all examples.

2.6 Application to a Glioblastoma Cancer Data

In this section, we apply our methodology to a glioblastoma multiforme (GBM) cancer data set studied by the Cancer Genome Atlas (TCGA) Research Network [TCGA, 2008]. As pointed out by TCGA, GBM is the most common primary form of brain tumor in adults. In our application, the data set contains 192 samples. Each sample has 11861 gene expression values and 535 microRNA expression values. Detailed documentation of the data

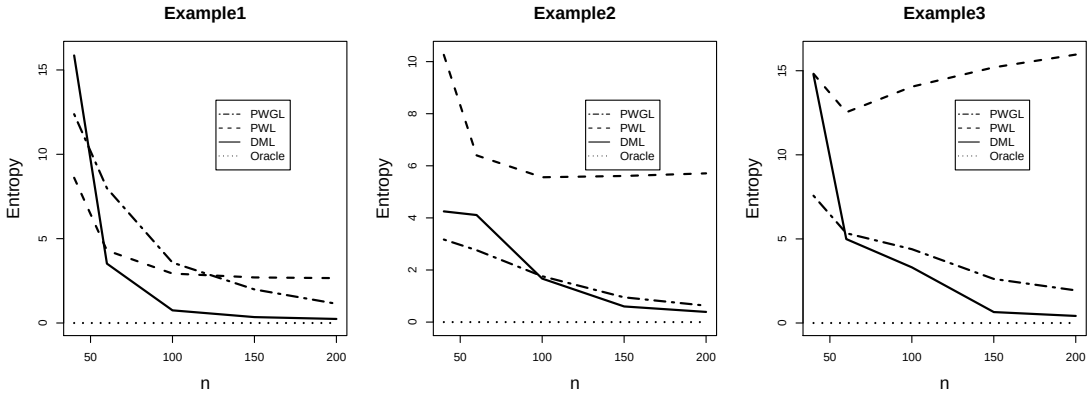

Figure 2.8: Averages of Entropy with p=40.

Table 2.3: Averages of PSE and the number of included genes based on 10 replications (The numbers in parentheses are standard errors).

	CW	RR	LASSO	PWL	DML
PSE	1.298 (0.038)	1.359 (0.045)	1.242 (0.035)	1.248 (0.032)	1.229 (0.032)
Number of included genes	500 (0.000)	500 (0.000)	158 (34.170)	17 (13.565)	78 (32.151)

can be found in Verhaak et al. [2010]. One of the main goals is to regress microRNAs on gene expressions to see how gene expressions can predict microRNAs. The other goal is to examine the underlying networks among microRNAs. We utilize the inverse covariance structure of microRNAs conditioned on gene expressions. To simplify the analysis, we perform prescreening to select a subset of genes. In particular, we use the median absolute deviation (MAD) to sort them and choose the top 500 genes with large MADs. Similar prescreening is also performed to choose top 20 microRNAs.

To examine performance of different methods, we divide this data set into training set, tuning set, and test set with 64 samples. Using the test set, we compare the performance of our methods with the other existing methods. The separate LASSO, the separate RR, and the CW are considered as competitors. Comparison among methods is performed in the two ways. First, we compare prediction accuracy by using PSE as the criterion. Second, we examine the number of included genes in each model.

Table 2.3 shows PSE and the number of included genes in each model. In terms of PSE, our method, DML, performs best even though the PSE difference between DML and the separate LASSO is not statistically significant in view of the standard errors. The PWL and the separate LASSO show similar performance while they outperform the separate RR

and the CW. In terms of the number of included genes, note that PWL and DML construct sparser models than the separate LASSO. One possible explanation of this is that there may be some strong positive correlations among microRNAs. As we have discussed in the toy example of Section 2.2, with strong positive correlations among response variables, joint methods tend to obtain more shrinkage than the separate LASSO. To explore this further, we examine correlations among the selected 20 microRNAs via scatter plots. Some strong correlations among the microRNAs are detected in scatter plots while negative correlations are not strong. Figure 2.9 shows some of the scatter plots. These scatter plots further demonstrate the usefulness of joint modeling and why our proposed PWL and DML methods obtain sparser models than the LASSO. Interestingly, with much fewer number of gene expressions than the separate LASSO, PWL and DML perform competitively in terms of prediction accuracy as shown in Table 2.3.

Figure 2.10 shows an estimated conditional inverse covariance structure of microRNAs given genes. The estimated inverse covariance is obtained from the model using our proposed DML method with one typical training set. On the left panel of Figure 2.10, two different microRNAs in Figure 2.9 are connected if their corresponding elements in the estimated inverse covariance is nonzero. We see that each pair of strongly correlated microRNAs in Figure 2.9 are also connected in this conditional dependence structure given genes. This may imply that the selected 500 genes have little effect on these microRNA correlations. As a final remark, we want to point out that the joint method may have numerical difficulty in very high dimensional problems as pointed out in Section 2.2.3. Prescreening can be very useful in that case.

2.7 Discussion

In this chapter, we proposed three methods for utilizing joint information among response variables in a penalized likelihood framework with weighted L_1 regularization. The proposed methods provide both sparse estimators for the regression parameter matrix and the conditional inverse covariance matrix of response vector given explanatory variables. Our theoretical investigation shows that our proposed estimators enjoys oracle properties. Simulated examples and an application to the GBM cancer data set demonstrate that our

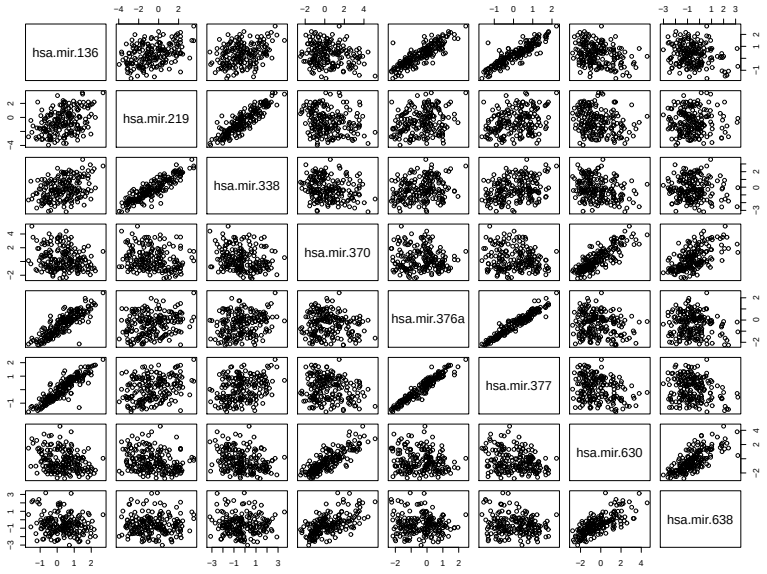

Figure 2.9: Scatter plots of eight selected microRNAs.

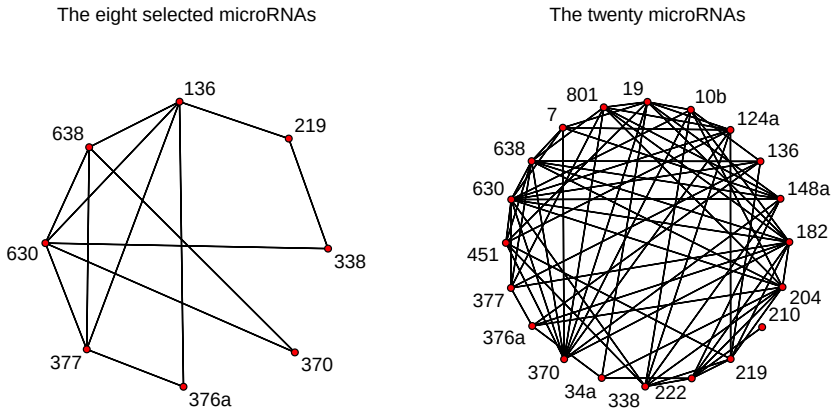

Figure 2.10: Graphical networks of the microRNAs using sparse inverse covariance structures of the microRNAs given the gene expressions.

proposed methods perform competitively.

Our current study assumes Gaussian distribution of the response vector. One future research direction is to extend the proposed method with other distribution assumptions. Although we mainly focus on the weighted L_1 penalty, our methods can be directly extended for other penalty functions as well. It will be interesting to compare the performance of various choices of penalty in this context.

2.8 Proofs

2.8.1 Proof of Lemma 1

Asymptotic normality

Let $\tilde{\mathbf{Y}} = ((\mathbf{y}^1)^T, ..., (\mathbf{y}^m)^T)^T$ be the nm-dimensional response vector and $\tilde{\boldsymbol{\epsilon}}$ be the corresponding nm-dimensional error vector which consists of $\epsilon_{ik}; i = 1, ..., n, k = 1, ..., m$. Let $\tilde{\boldsymbol{\beta}} = (\beta_{11}, ..., \beta_{p1}, ..., \beta_{1m}, ..., \beta_{pm})^T$ be the pm-dimensional vector and $\tilde{\mathbf{X}} = \mathbf{I}_m \otimes \mathbf{X}$. Then the minimizer of (2.6) is equivalent to

$$\underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} \left[(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}})^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}) + \lambda_1 \sum_{j,k} w_{jk} \left|\beta_{jk}\right| \right].$$

Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}$ and

$$V_n(\mathbf{u}) = (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}))^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}})) + \lambda_{1,n} \sum_{j,k} w_{jk} |\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}|.$$

Let $\hat{\mathbf{u}}^{(n)} = \operatorname{argmin}_{\mathbf{u}} V_n(\mathbf{u})$ and then $\hat{\mathbf{u}}^{(n)} = \sqrt{n}(\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)} - \boldsymbol{\beta}^*)$. Note that $\hat{\mathbf{u}}^{(n)} = \operatorname{argmin}_{\mathbf{u}} V_n(\mathbf{u}) = \operatorname{argmin}_{\mathbf{u}} \{V_n(\mathbf{u}) - V_n(\mathbf{0})\}$ and

$$\begin{aligned} V_n(\mathbf{u}) - V_n(\mathbf{0}) = \frac{1}{n}\mathbf{u}^T \tilde{\mathbf{X}}^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u} - \frac{2}{\sqrt{n}}\tilde{\boldsymbol{\epsilon}}^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u} \\ + \lambda_{1,n} \sum_{j,k} w_{jk} \left(\left| \beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}} \right| - |\beta^*_{jk}| \right). \end{aligned} \tag{2.17}$$

We know that $\frac{1}{n}\mathbf{u}^T \tilde{\mathbf{X}}^T (\boldsymbol{\Omega} \otimes I_n)\tilde{\mathbf{X}}\mathbf{u} = \mathbf{u}^T (\boldsymbol{\Omega} \otimes \frac{1}{n}\mathbf{X}^T\mathbf{X})\mathbf{u} \to \mathbf{u}^T (\boldsymbol{\Omega} \otimes A)\mathbf{u}$. For the second term of the right hand side of (2.17), note that $\tilde{\boldsymbol{\epsilon}} \sim \mathbf{N}(0, \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$. Thus, $\frac{1}{\sqrt{n}}\tilde{\boldsymbol{\epsilon}}^T (\boldsymbol{\Omega} \otimes I_n)\tilde{\mathbf{X}} \to_d \mathbf{Z}$ where $\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\Omega} \otimes A)$ as $\frac{1}{n}\tilde{\mathbf{X}}^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}} = \frac{1}{n}\tilde{\mathbf{X}}^T (\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}} \to \boldsymbol{\Omega} \otimes A$. Now

we consider the last term of the right hand side of (2.17):

- If $\beta^*_{jk} = 0$, then $\lambda_{1,n} w_{jk}(\left|\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - \left|\beta^*_{jk}\right|) = \frac{\lambda_{1,n}}{\sqrt{n}} w_{jk}|u_{jk}| = \lambda_{1,n} n^{\frac{\gamma-1}{2}} \frac{|u_{jk}|}{(\sqrt{n}|\tilde{\beta}_{jk}|)^\gamma} \to \infty$ as $\sqrt{n}\tilde{\beta}_{jk} = O_p(1)$.

- if $\beta^*_{jk} \neq 0$, then $\lambda_{1,n} w_{jk}(\left|\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - \left|\beta^*_{jk}\right|) = \frac{\lambda_{1,n}}{\sqrt{n}} w_{jk} \sqrt{n}(\left|\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - \left|\beta^*_{jk}\right|)$. Note that $\frac{\lambda_{1,n}}{\sqrt{n}} \to 0$, $w_{jk} \to_p \frac{1}{|\beta^*_{jk}|^\gamma}$ and $\sqrt{n}(\left|\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - \left|\beta^*_{jk}\right|) \to u_{jk}\text{sign}(\beta^*_{jk})$. By the Slutsky's theorem, $\lambda_{1,n} w_{jk}(\left|\beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - \left|\beta^*_{jk}\right|) \to_p 0$.

By combining above statements and using the Slutsky's theorem again, we obtain the following:

$$V_n(\mathbf{u}) - V_n(\mathbf{0}) \to_d V(\mathbf{u}) = \begin{cases} \mathbf{u}_{\mathcal{A}}^T D \mathbf{u}_{\mathcal{A}} - 2\mathbf{u}_{\mathcal{A}}^T \mathbf{Z}_{\mathcal{A}} & \text{if} \quad u_{jk} = 0 \quad \text{for all} \quad (j,k) \notin \mathcal{A}, \\ \infty & \text{if} \quad \text{otherwise}, \end{cases}$$

where $\mathbf{u}_{\mathcal{A}}$ consists of u_{jk} for $(j,k) \in \mathcal{A}$ and $\mathbf{Z}_{\mathcal{A}} \sim \mathbf{N}(0, D)$.

Let $\hat{\mathbf{u}} = \text{argmin}_{\mathbf{u}} V(\mathbf{u})$. Then we have

$$\begin{cases} \hat{\mathbf{u}}_{\mathcal{A}} = D^{-1}\mathbf{Z}_{\mathcal{A}}, \\ \hat{u}_{jk} = 0 \quad \forall (j,k) \notin \mathcal{A}. \end{cases}$$

Note that $V_n(\mathbf{u}) - V_n(\mathbf{0})$ is convex and so $\text{argmin}_{\mathbf{u}}(V_n(\mathbf{u}) - V_n(\mathbf{0})) \to_d \text{argmin}_{\mathbf{u}} V(\mathbf{u})$. Since $\mathbf{Z}_{\mathcal{A}} \sim \mathbf{N}(0, D)$, thus $\hat{\mathbf{u}}^{(n)}_{\mathcal{A}} \to_d \mathbf{N}(0, D^{-1})$. Finally, we have that $\hat{\mathbf{u}}^{(n)}_{\mathcal{A}} = \sqrt{n}(\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d D^{-1}\mathbf{Z}_{\mathcal{A}}$ as $n \to \infty$.

Selection consistency

We need to show that $\forall (j,k) \notin \mathcal{A}$, $P(\hat{\beta^1}^{(n)}_{jk} \neq 0) \to 0$. For fixed $(j,k) \notin \mathcal{A}$, let $(j,k) \in \mathcal{A}^1_n$. Then $|\hat{\beta^1}^{(n)}_{jk}| \neq 0$ and so we have that $2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)}) = \lambda_{1,n} w_{jk} \text{sign}(\hat{\beta^1}^{(n)}_{jk})$ by the KKT conditions where $\tilde{\mathbf{x}}_{jk}$ is $(j + (k-1))$-th row of $\tilde{\mathbf{X}}$. Therefore, $P(\hat{\beta^1}^{(n)}_{jk} \neq 0) \leq P(2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)}) = \lambda_{1,n} w_{jk} \text{sign}(\hat{\beta^1}^{(n)}_{jk}))$. Note that

$$\frac{2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)})}{\sqrt{n}} = \frac{2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)})}{n} + \frac{2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\boldsymbol{\epsilon}}}{\sqrt{n}}.$$

From the asymptotic normality part, we know that $\frac{2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)})}{n}$ converges in distribution to some normal random vector. We also have that $\frac{2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)\tilde{\boldsymbol{\epsilon}}}{\sqrt{n}} \to_d \mathbf{N}(0, (\boldsymbol{\Omega} \otimes A)_{jk,jk})$ where $(\boldsymbol{\Omega} \otimes A)_{jk,jk}$ is the $(j+(k-1))$-th diagonal element of $\boldsymbol{\Omega} \otimes A$. As $\frac{\lambda_{1,n} w_{jk} \text{sign}(\hat{\beta^1}^{(n)}_{jk})}{\sqrt{n}} = \lambda_{1,n} n^{\frac{\gamma-1}{2}} \frac{\text{sign}(\hat{\beta^1}^{(n)}_{jk})}{(\sqrt{n}|\tilde{\beta}_{jk}|)^\gamma} \to \pm\infty$ with $\sqrt{n}\tilde{\beta}_{jk} = O_p(1)$, we have $P(2\tilde{\mathbf{x}}^T_{jk}(\boldsymbol{\Omega} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)}) =$

$\lambda_{1,n} w_{jk} \text{sign}(\hat{\beta^1}_{jk}^{(n)})) \to 0$. Therefore, $P(\hat{\beta^1}_{jk}^{(n)} \neq 0) \to 0$ as $n \to \infty$.

2.8.2 Proof of Theorem 1

The proof is similar to that of Lemma 1 except we replace $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}}$.

Asymptotic normality

Note that (2.8) is equivalent to

$$\underset{\tilde{\boldsymbol{\beta}}}{\text{argmin}} \left[(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}})^T (\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}) + \lambda_1 \sum_{j,k} w_{jk} \left|\beta_{jk}\right| \right].$$

Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}$ and

$$V_n^*(\mathbf{u}) = (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}))^T (\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}})) + \lambda_{1,n} \sum_{j,k} w_{jk} \left| \beta^*_{jk} + \frac{u_{jk}}{\sqrt{n}} \right|.$$

Let $\hat{\mathbf{u}}^{(n)} = \text{argmin}_{\mathbf{u}} V_n^*(\mathbf{u})$ and then $\hat{\mathbf{u}}^{(n)} = \sqrt{n}(\hat{\boldsymbol{\beta}^{\mathbf{2}}}^{(n)} - \boldsymbol{\beta}^*)$. We can show that

$$\begin{aligned} V_n^*(\mathbf{u}) - V_n^*(\mathbf{0}) =& V_n(\mathbf{u}) - V_n(\mathbf{0}) + \frac{1}{n}\mathbf{u}^T \tilde{\mathbf{X}}^T ((\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u} \\ & - \frac{2}{\sqrt{n}} \tilde{\boldsymbol{\epsilon}}^T ((\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u}, \end{aligned}$$

where $V_n(\mathbf{u})$ is defined in the proof of Lemma 1. As the $\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\boldsymbol{\Omega}$, $\frac{1}{n}\mathbf{u}^T \tilde{\mathbf{X}}^T ((\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u} \to_p 0$ and $\frac{2}{\sqrt{n}} \tilde{\boldsymbol{\epsilon}}^T ((\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u} \to_d 0$. From the proof of Lemma 1, we also know that $V_n(\mathbf{u}) - V_n(\mathbf{0}) \to_d V(\mathbf{u})$. By combining the above statements and using Slutsky's theorem, we have that $V_n^*(\mathbf{u}) - V_n^*(\mathbf{0}) \to_d V(\mathbf{u})$. By using the same arguments in the proof of Lemma 1, finally we have that $\hat{\mathbf{u}}_{\mathcal{A}}^{(n)} = \sqrt{n}(\hat{\boldsymbol{\beta}^{\mathbf{2}}}_{\mathcal{A}}^{(n)} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d D^{-1}\mathbf{Z}_{\mathcal{A}}$ as $n \to \infty$.

Selection consistency

Now it suffices to show that $\forall (j,k) \notin \mathcal{A}$, $P(\hat{\beta^2}_{jk}^{(n)} \neq 0) \to 0$ as $n \to \infty$. For fixed $(j,k) \notin \mathcal{A}$, let $(j,k) \in \mathcal{A}_n^2$. Then $|\hat{\beta^2}_{jk}^{(n)}| \neq 0$ and so we have that $2\tilde{\mathbf{x}}_{jk}^T (\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}^{\mathbf{2}}}^{(n)}) = \lambda_{1,n} w_{jk} \text{sign}(\hat{\beta^2}_{jk}^{(n)})$ by the KKT conditions. Therefore, $P(\hat{\beta^2}_{jk}^{(n)} \neq 0) \leq P(2\tilde{\mathbf{x}}_{jk}^T (\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} -$

$\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)}) = \lambda_{1,n} w_{jk} \text{sign}(\hat{\beta}^{2(n)}_{jk}))$. Note that

$$\frac{2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)})}{\sqrt{n}} = \frac{2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathbf{2}(n)})}{n} + \frac{2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)\tilde{\boldsymbol{\epsilon}}}{\sqrt{n}}.$$

From the asymptotic normality part and the fact that $\hat{\boldsymbol{\Omega}}$ is consistent, we know that $\frac{2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)\tilde{\mathbf{X}}\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathbf{2}(n)})}{n}$ converges in distribution to some normal random vector. We also have that $\frac{2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)\tilde{\boldsymbol{\epsilon}}}{\sqrt{n}} \to_d \mathbf{N}(0, (\boldsymbol{\Omega} \otimes A)_{jk,jk})$. As $\frac{\lambda_{1,n} w_{jk} \text{sign}(\hat{\beta}^{2(n)}_{jk})}{\sqrt{n}} = \lambda_{1,n} n^{\frac{\gamma-1}{2}} \frac{\text{sign}(\hat{\beta}^{2(n)}_{jk})}{(\sqrt{n}|\hat{\beta}_{jk}|)^\gamma} \to \pm\infty$, we have $P(2\tilde{\mathbf{x}}_{jk}^T(\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_n)(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^{\mathbf{2}(n)}) = \lambda_{1,n} w_{jk} \text{sign}(\hat{\beta}^{2(n)}_{jk})) \to 0$. Therefore, $P(\hat{\beta}^{2(n)}_{jk} \neq 0) \to 0$ as $n \to \infty$.

2.8.3 Proof of Lemma 2

Let $R = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)$. With given $\mathbf{B}^*$, define $Q(\boldsymbol{\Omega})$ as

$$Q(\boldsymbol{\Omega}) = -n \log\det(\boldsymbol{\Omega}) + n\text{tr}(\boldsymbol{\Omega}R) + \lambda_{2,n} \sum_{j \neq k} v_{jk}|\omega_{jk}|. \tag{2.18}$$

Selection consistency

Using the definition of $Q(\boldsymbol{\Omega})$ in (2.18), define $V_n(U)$ as

$$\begin{aligned} V_n(U) &= Q(\boldsymbol{\Omega}^* + \frac{U}{\sqrt{n}}) - Q(\boldsymbol{\Omega}^*) \\ &= -n \log\det((\boldsymbol{\Omega}^* + \frac{U}{\sqrt{n}})\boldsymbol{\Omega}^{*-1}) + n\text{tr}(\frac{UR}{\sqrt{n}}) + \lambda_{2,n} \sum_{j \neq k} v_{jk}(|\omega^*_{jk} + \frac{u_{jk}}{\sqrt{n}}| - |\omega^*_{jk}|). \end{aligned}$$

Using a similar argument as in the proof of Theorem 1 in Yuan and Lin (2007), it can be shown that

$$V_n(U) = \text{tr}(U\boldsymbol{\Sigma}U\boldsymbol{\Sigma}) + \text{tr}[U\sqrt{n}(R - \boldsymbol{\Sigma})] + \lambda_{2,n} \sum_{j \neq k} v_{jk}(|\omega^*_{jk} + \frac{u_{jk}}{\sqrt{n}}| - |\omega^*_{jk}|) + o(1).$$

Note that as $v_{st} = \frac{1}{|\tilde{\omega}_{st}|}$, $\lambda_{2,n} n^{-\frac{1}{2}} \to 0$, and $\tilde{\omega}_{jk} \to_p \omega^*_{jk}$, we have

$$\lambda_{2,n} \sum_{j \neq k} v_{jk}(|\omega^*_{jk} + \frac{u_{jk}}{\sqrt{n}}| - |\omega^*_{jk}|) = \lambda_{2,n} \sum_{\omega^*_{jk}=0} \frac{|u_{jk}|}{\sqrt{n}|\tilde{\omega}_{jk}|} + \frac{\lambda_{2,n}}{\sqrt{n}} \sum_{\omega^*_{jk} \neq 0} (\frac{|u_{jk}|}{|\tilde{\omega}_{jk}|} \text{sign}(\omega^*_{jk}) + o(1))$$
$$= \lambda_{2,n} \sum_{\omega^*_{jk}=0} \frac{|u_{jk}|}{\sqrt{n}|\tilde{\omega}_{jk}|} + o_p(1).$$

On the other hand, $\sqrt{n}(R - \mathbf{\Sigma}) \to_d \mathbf{N}(0, \Lambda)$ by the central limit theorem as $R = \frac{1}{n} \sum_i^n \epsilon_i \epsilon_i^T$. Therefore, $V_n(U)$ can be written as

$$V_n(U) = \text{tr}(U\mathbf{\Sigma}U\mathbf{\Sigma}) + \text{tr}(UW_n) + \lambda_{2,n} \sum_{\omega^*_{jk}=0} \frac{|u_{jk}|}{\sqrt{n}|\tilde{\omega}_{jk}|} + o_p(1),$$

where $W_n \to_d \mathbf{N}(0, \Lambda)$. Denote by $\hat{U}$ the minimizer of $V_n(U)$. Note that $\lambda_{2,n} \to \infty$ and $\sqrt{n}|\tilde{\omega}_{jk}| = O_p(1)$. Therefore, if $\omega^*_{jk} = 0$, $P(\hat{u}_{jk} = 0) \to 1$ as $n \to \infty$. This completes the proof of the variable selection consistency.

Asymptotic normality

Suppose U satisfies that $u_{jk} = 0$ if $\omega^*_{jk} = 0$. Then, $V_n(U)$ can be written as

$$V_n(U) = \text{tr}(U\mathbf{\Sigma}U\mathbf{\Sigma}) + \text{tr}[U\sqrt{n}(R - \mathbf{\Sigma})] + o_p(1).$$

By using the Slutsky's theorem, we have that

$$V_n(U) \to_d V(U) = \text{tr}(U\mathbf{\Sigma}U\mathbf{\Sigma}) + \text{tr}(UW) \quad \text{where} \quad \text{vec}(W) \sim \mathbf{N}(0, \Lambda).$$

Since $V_n(U)$ and $V(U)$ are both convex and $V(U)$ has a unique minimum, $\text{argmin}\, V_n(U) \to_d \text{argmin}\, V(U)$. From the fact that $\text{argmin}\, V_n(U) = \text{argmin}\, Q(\mathbf{\Omega}^* + \frac{U}{\sqrt{n}}) = \sqrt{n}(\hat{\mathbf{\Omega}}_0^1 - \mathbf{\Omega}^*)$, $\text{argmin}\, V_n(U) = \sqrt{n}(\hat{\mathbf{\Omega}}_0^1 - \mathbf{\Omega}^*) \to_d \text{argmin}\, V(U)$. This completes the proof of the asymptotic normality.

2.8.4 Proof of Theorem 2

With a $\sqrt{n}$-consistent estimator $\hat{\mathbf{B}}$ of $\mathbf{B}$, let $\hat{R} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$. Define $Q(\mathbf{\Omega})$ as

$$Q(\mathbf{\Omega}) = -n\log\det(\mathbf{\Omega}) + n\mathrm{tr}(\mathbf{\Omega}\hat{R}) + \lambda_{2,n}\sum_{j\neq k} v_{jk}|\omega_{jk}|. \tag{2.19}$$

By using the above definition, define $V_n(U)$ as

$$\begin{aligned} V_n(U) &= Q(\mathbf{\Omega}^* + \frac{U}{\sqrt{n}}) - Q(\mathbf{\Omega}^*) \\ &= -n\log\det((\mathbf{\Omega}^* + \frac{U}{\sqrt{n}})\mathbf{\Omega}^{*-1}) + n\mathrm{tr}(\frac{U\hat{R}}{\sqrt{n}}) + \lambda_{2,n}\sum_{j\neq k} v_{jk}(|\omega^*_{jk} + \frac{u_{jk}}{\sqrt{n}}| - |\omega^*_{jk}|). \end{aligned}$$

Note that

$$n\mathrm{tr}(\frac{U\hat{R}}{\sqrt{n}}) = n\mathrm{tr}(\frac{U(\hat{R}-R)}{\sqrt{n}}) + n\mathrm{tr}(\frac{UR}{\sqrt{n}}).$$

Therefore, by the proof of Lemma 2 and the Slutsky's theorem, it suffices to show that

$$n\mathrm{tr}(\frac{U(\hat{R}-R)}{\sqrt{n}}) = o_p(1). \tag{2.20}$$

The left-hand side of (2.20) can be written as

$$\begin{aligned} n\mathrm{tr}(\frac{U(\hat{R}-R)}{\sqrt{n}}) &= \mathrm{tr}(\frac{U}{\sqrt{n}}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})) - \mathrm{tr}(\frac{U}{\sqrt{n}}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})) \\ &= \mathrm{tr}(U\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B})^T\frac{\mathbf{X}^T\mathbf{X}}{n}(\hat{\mathbf{B}} - \mathbf{B})) - 2\mathrm{tr}(U\frac{(\mathbf{Y} - \mathbf{X}\mathbf{B})^T\mathbf{X}}{\sqrt{n}}(\hat{\mathbf{B}} - \mathbf{B})), \end{aligned}$$

where we add and subtract $\mathbf{XB}$ in the first term. Since $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = O_p(1)$, $\frac{(\mathbf{Y}-\mathbf{X}\mathbf{B})^T\mathbf{X}}{\sqrt{n}} = O_p(1)$, $(\hat{\mathbf{B}} - \mathbf{B}) = o_p(1)$ and $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to A$, (2.20) holds.

2.8.5 Proof of Lemma 3

Define $Q(\mathbf{B}, \mathbf{\Omega})$ for the jointly penalized likelihood as

$$Q(\mathbf{B}, \mathbf{\Omega}) = -n\log\det(\mathbf{\Omega}) + \mathrm{tr}\left\{\mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\right\} + \lambda_{1,n}\sum_{j,k} w_{jk}\left|\beta_{jk}\right| + \lambda_{2,n}\sum_{s\neq t} v_{st}\left|\omega_{st}\right|. \tag{2.21}$$

To show the results, we use the similar idea of the proof of Theorem 1 in Fan and Li [2001]. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\{\sup_{\|U\|=D} Q(\mathbf{B}^* + \frac{U_1}{\sqrt{n}}, \mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}}) > Q(\mathbf{B}^*, \mathbf{\Omega}^*)\} \geq 1 - \delta, \tag{2.22}$$

where $U = (\mathrm{vec}(U_1)^T, \mathrm{vec}(U_2)^T)^T$. Using the definition of $Q(\mathbf{B}, \mathbf{\Omega})$ in (2.21), define $V_n(U)$ as

$$V_n(U) = Q(\mathbf{B}^* + \frac{U_1}{\sqrt{n}}, \mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}}) - Q(\mathbf{B}^*, \mathbf{\Omega}^*).$$

Since $|\beta^*_{jk} + \frac{u_{1jk}}{\sqrt{n}}| - |\beta^*_{jk}| = |\frac{u_{1jk}}{\sqrt{n}}|$ for $\beta^*_{jk} = 0$ and $|\omega^*_{st} + \frac{u_{2st}}{\sqrt{n}}| - |\omega^*_{st}| = |\frac{u_{2st}}{\sqrt{n}}|$ for $\omega^*_{st} = 0$,

$$\begin{aligned}
V_n(U) \geq & -n \log\det((\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})\mathbf{\Omega}^{*-1}) + \mathrm{tr}\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})(\mathbf{Y} - \mathbf{X}(\mathbf{B}^* + \frac{U_1}{\sqrt{n}}))^T(\mathbf{Y} - \mathbf{X}(\mathbf{B}^* + \frac{U_1}{\sqrt{n}}))\} \\
& - \mathrm{tr}\{\mathbf{\Omega}^*(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)\} + \lambda_{1,n} \sum_{\beta_{kj} \neq 0} w_{jk}(|\beta^*_{jk} + \frac{u_{1jk}}{\sqrt{n}}| - |\beta^*_{jk}|) \\
& + \lambda_{2,n} \sum_{\omega_{st} \neq 0} v_{st}(|\omega^*_{st} + \frac{u_{2st}}{\sqrt{n}}| - |\omega^*_{st}|) \\
= & -n \log\det((\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})\mathbf{\Omega}^{*-1}) + \mathrm{tr}\{\frac{U_2}{\sqrt{n}}(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)\} \\
& + \mathrm{tr}\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})(\frac{\mathbf{X}U_1}{\sqrt{n}})^T(\frac{\mathbf{X}U_1}{\sqrt{n}})\} - 2\mathrm{tr}\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\frac{\mathbf{X}U_1}{\sqrt{n}})\} \\
& + \lambda_{1,n} \sum_{\beta_{kj} \neq 0} w_{jk}(|\beta^*_{jk} + \frac{u_{1jk}}{\sqrt{n}}| - |\beta^*_{jk}|) + \lambda_{2,n} \sum_{\omega_{st} \neq 0} v_{st}(|\omega^*_{st} + \frac{u_{2st}}{\sqrt{n}}| - |\omega^*_{st}|).
\end{aligned} \tag{2.23}$$

For the first term and the second term on the right-hand side of (2.23), it has been shown in Lemma 2 that

$$-n \log\det((\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})\mathbf{\Omega}^{*-1}) + \mathrm{tr}\{\frac{U_2}{\sqrt{n}}(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)\} = \mathrm{tr}(U_2\mathbf{\Sigma}U_2\mathbf{\Sigma}) + \mathrm{tr}(U_2 W_n).$$

Let $\tilde{U}_1 = \mathrm{vec}(U_1)$. For the third term on the right-hand side of (2.23), as $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to A$, note that

$$\mathrm{tr}\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})(\frac{\mathbf{X}U_1}{\sqrt{n}})^T(\frac{\mathbf{X}U_1}{\sqrt{n}})\} = \tilde{U}_1^T\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}}) \otimes (\frac{\mathbf{X}^T\mathbf{X}}{n})\}\tilde{U}_1 = \tilde{U}_1^T(\mathbf{\Omega}^* \otimes A)\tilde{U}_1 + o(1).$$

For the fourth term on the right-hand side of (2.23), we have

$$\mathrm{tr}\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}})(\mathbf{Y} - \mathbf{X}\mathbf{B}^*)^T(\frac{\mathbf{X}U_1}{\sqrt{n}})\} = \tilde{U}_1^T(\frac{\tilde{\mathbf{X}}}{\sqrt{n}})^T\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}}) \otimes \mathbf{I}_n\}\tilde{\boldsymbol{\epsilon}}.$$

Note that $(\frac{\tilde{\mathbf{X}}}{\sqrt{n}})^T\{(\mathbf{\Omega}^* + \frac{U_2}{\sqrt{n}}) \otimes \mathbf{I}_n\}\tilde{\boldsymbol{\epsilon}} \to_d Z$ where Z has multivariate normal distribution of dimension $n \times m$. By combining above statements, we have

$$\begin{aligned} V_n(U) \geq\ & \mathrm{tr}(U_2\mathbf{\Sigma}U_2\mathbf{\Sigma}) + \mathrm{tr}(U_2W_n) + \tilde{U}_1^T(\mathbf{\Omega}^* \otimes A)\tilde{U}_1 + \tilde{U}_1^T Z_n + o_p(1) \\ & + \lambda_{1,n}\sum_{\beta_{kj}\neq 0} w_{jk}(|\beta_{jk}^* + \frac{u_{1jk}}{\sqrt{n}}| - |\beta_{jk}^*|) + \lambda_{2,n}\sum_{\omega_{st}\neq 0} v_{st}(|\omega_{st}^* + \frac{u_{2st}}{\sqrt{n}}| - |\omega_{st}^*|). \end{aligned}$$

As $\lambda_{1,n}n^{-\frac{1}{2}} \to 0$ and $\lambda_{2,n}n^{-\frac{1}{2}} \to 0$, we have

$$\begin{aligned} \lambda_{1,n}\sum_{\beta_{jk}^*\neq 0} w_{jk}(|\beta_{jk}^* + \frac{u_{1jk}}{\sqrt{n}}| - |\beta_{jk}^*|) &= \frac{\lambda_{1,n}}{\sqrt{n}}\sum_{\beta_{jk}^*\neq 0}(\frac{|u_{1jk}|}{|\tilde{\beta}_{jk}|^\gamma}\mathrm{sign}(\beta_{jk}^*) + o(1)) = o_p(1), \\ \lambda_{2,n}\sum_{\omega_{st}\neq 0} v_{st}(|\omega_{st}^* + \frac{u_{2st}}{\sqrt{n}}| - |\omega_{st}^*|) &= \frac{\lambda_{2,n}}{\sqrt{n}}\sum_{\omega_{st}^*\neq 0}(\frac{|u_{2st}|}{|\tilde{\omega}_{st}|}\mathrm{sign}(\omega_{st}^*) + o(1)) = o_p(1). \end{aligned}$$

Therefore,

$$V_n(U) \geq \mathrm{tr}(U_2\mathbf{\Sigma}U_2\mathbf{\Sigma}) + \mathrm{tr}(U_2W_n) + \tilde{U}_1^T(\mathbf{\Omega}^* \otimes A)\tilde{U}_1 + \tilde{U}_1^T Z_n + o_p(1). \tag{2.24}$$

By choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \|U\| = D\}$ with the probability greater than $1 - \delta$ as $\mathbf{\Omega}^*$ and A are positive-definite, $W_n = O_p(1)$, and $Z_n = O_p(1)$. Therefore, (2.22) holds. This completes the proof of this lemma.

2.8.6 Proof of Theorem 3

As defined in Lemma 3, define $Q(\mathbf{B}, \mathbf{\Omega})$ for the jointly penalized likelihood as

$$Q(\mathbf{B}, \mathbf{\Omega}) = -n\log\det(\mathbf{\Omega}) + \mathrm{tr}\left\{\mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\right\} + \lambda_{1,n}\sum_{j,k} w_{jk}\left|\beta_{jk}\right| + \lambda_{2,n}\sum_{s\neq t} v_{st}\left|\omega_{st}\right|.$$

Note that $(\hat{\mathbf{B}}^{(n)}, \hat{\mathbf{\Omega}})$ is a $\sqrt{n}$-consistent local minimizer of $Q(\mathbf{B}, \mathbf{\Omega})$. As $\hat{\mathbf{B}}^{(n)} = \mathrm{argmin}_{\mathbf{B}}\, Q(\mathbf{B}, \hat{\mathbf{\Omega}})$ and $\hat{\mathbf{\Omega}}$ is $\sqrt{n}$-consistent, the oracle properties of $\hat{\mathbf{B}}^{(n)}$ hold by Theorem 1. Similarly, since $\hat{\mathbf{\Omega}} = \mathrm{argmin}_{\mathbf{\Omega}}\, Q(\hat{\mathbf{B}}^{(n)}, \mathbf{\Omega})$ and $\hat{\mathbf{B}}^{(n)}$ is $\sqrt{n}$-consistent, the oracle properties of $\hat{\mathbf{\Omega}}$ hold by

Theorem 2. These complete the proof of this theorem.

Chapter 3

Multiple Response Regression with Mixture Gaussian Models

3.1 Introduction

In Chapter 2, we considered the multivariate response regression problem under the assumption of multivariate Gaussian distribution. In particular, we assume that given the covariates, the response vector follows an m-dimensional Gaussian distribution. In Section 2.2, we considered the three methods, PWL method, PWGL method, and DML method.

The methods in Chapter 2 can be very useful in dealing with high dimensional multivariate Gaussian data. However, in some applications, the assumption of a single Gaussian distribution can be too strong. For example, in the GBM dataset considered in Section 2.6, Verhaak et al. [2010] showed that the GBM patients can be divided into four subtypes based on their gene expressions and they call name as subtypes of Classical, Mesenchymal, Neural, and Proneural. The gene expressions of patients within each subtype can be very similar as shown in Figure 3.1. However, the gene expressions of 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. For instance, Figure 3.2 shows the histogram and densities of gene expression levels of *EGFR* which is known to be associated with the GBM cancer [Verhaak et al., 2010]. In Figure 3.2, we can see that there are multiple modes in the distribution of all observations, which is not appropriate for a single multivariate Gaussian distribution. On the other hand, within each subtype, the corresponding distribution of the expression levels appears more reasonable for a normal

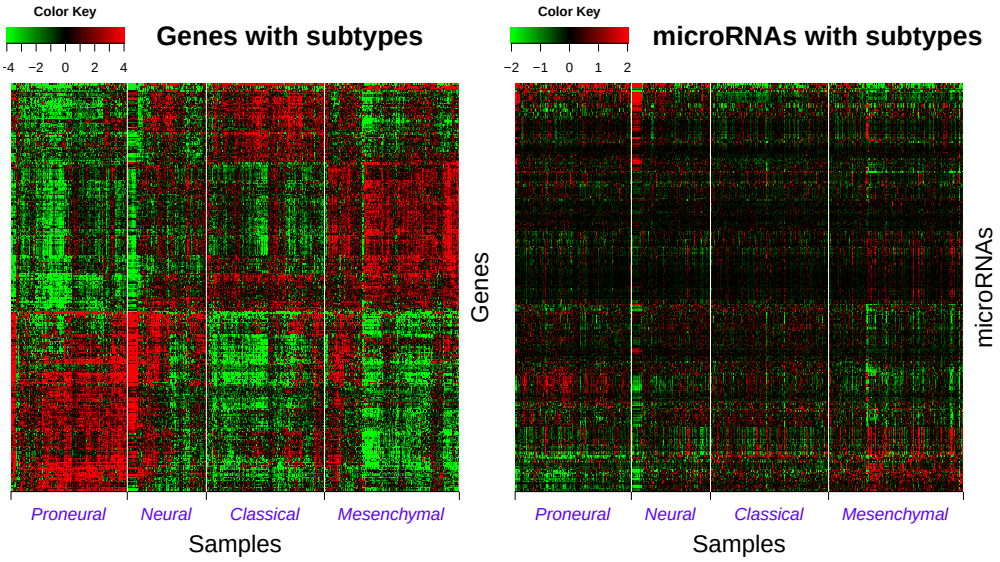

Figure 3.1: Heatmaps of gene and microRNA expressions of GBM patients with four subtypes.

distribution. The main reason for this phenomenon is that *EGFR* tends to be highly expressed in the classical subtype. In this chapter, 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. A naive approach to tackle this problem is to model each group separately. However, this approach ignores the common structure that may exist across different groups. Therefore, it might be more useful to model all groups jointly so that the common structure can be estimated from the aggregated data.

In this chapter, parallel to the methods in Chapter 2, 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 parameter matrices. For the second approach, we estimate the inverse covariance matrices instead after plugging in a good estimator of the regression parameter matrices. The last approach simultaneously estimates the regression parameter matrices and the inverse covariance matrices. These methods are penalized log-likelihood approaches with the multivariate mixture Gaussian assumption.

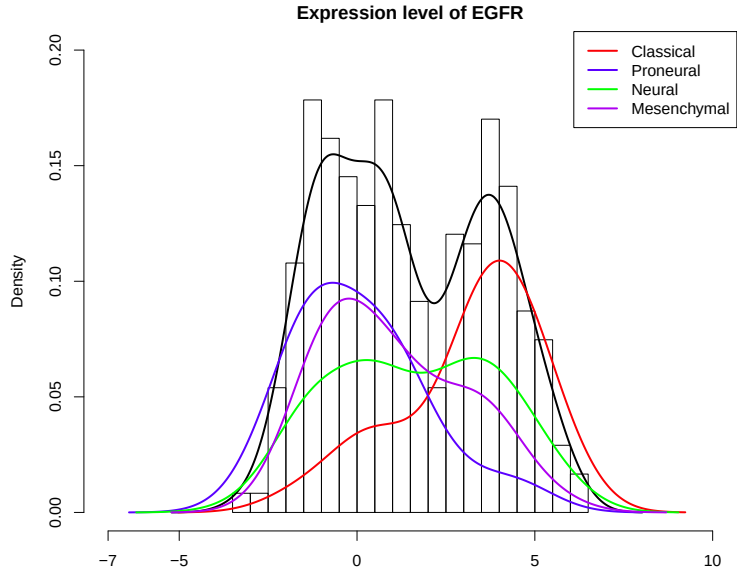

Figure 3.2: Expression levels of the gene *EGFR* with four subtypes.

In the following sections, we describe the new proposed methods in more details with theoretical justification. In Section 3.2, we introduce our proposed methodology. Section 3.3 explores theoretical properties of our proposed methods. Section 3.4 develops computational algorithms to obtain solutions for proposed methods. Simulated examples are presented in Section 3.5 to demonstrate performance of our methods and Section 3.6 provides analysis of the glioblastoma cancer data. Section 3.7 provides some discussions. The proofs of the theorems are provided in Section 3.8.

3.2 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 $\mathbf{y}_i^{(g)} = (y_{i1}^{(g)}, ..., y_{im}^{(g)})^T; i = 1, ..., n_g$, be m-dimensional responses and $\mathbf{Y}^{(g)} = [\mathbf{y}_1^{(g)}, ..., \mathbf{y}_{n_g}^{(g)}]^T$ be the $n_g \times m$ response matrix in the g-th group. Let $\mathbf{x}_i^{(g)} = (x_{i1}^{(g)}, ..., x_{ip}^{(g)})^T; i = 1, ..., n_g$, be p-dimensional predictors and $\mathbf{X}^{(g)} = [\mathbf{x}_1^{(g)}, ..., \mathbf{x}_{n_g}^{(g)}]^T$ 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,

$$\mathbf{Y}^{(g)} = \mathbf{X}^{(g)}\mathbf{B}^{(g)} + \mathbf{e}^{(g)}, \quad \text{with} \quad \mathbf{e}^{(g)} = [\epsilon_1^{(g)}, ..., \epsilon_{n_g}^{(g)}]^T,$$

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

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

$$\frac{n_g}{2}\log\det(\mathbf{\Omega}^{(g)}) - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{\Omega}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}, \quad g = 1, ..., G, \tag{3.1}$$

up to a constant not depending on $(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})$. As stated in Section 2.2, 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 estimation procedure, in Section 2.2.3, we proposed the DML method. The estimator is given by solving

$$\underset{\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)}}{\mathrm{argmin}} \left\{-l(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)}) + \lambda_1 \sum_{j,k}\left|\beta_{jk}^{(g)}\right| + \lambda_2 \sum_{s\neq t}\left|\omega_{st}^{(g)}\right|\right\}. \tag{3.2}$$

where $l(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)}) = n_g \log\det(\mathbf{\Omega}^{(g)}) - \mathrm{tr}\left\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{\Omega}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}; g = 1, ..., G$. By using L_1 penalties, the joint information, $\mathbf{\Omega}^{(g)}$, has an effect on the estimation of $\mathbf{B}^{(g)}$.

Motivated from the technique for a single linear model as in (3.2), we consider penalization for (3.1) to improve estimation. In particular, estimation of $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$ can be improved if some common information across groups can be shared in the estimation procedure. Note that the optimization problem suggested in (3.2) 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 $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$ simultaneously to identify the

common and unique structures across groups. Note that there are two parameter matrices in each group, $\mathbf{B}^{(g)}$ and $\mathbf{\Omega}^{(g)}$, involved in the estimation. It is also often that only one of them is of main interest. Hence, parallel to the methods in Chapter 2, we consider three different approaches, two plug-in methods and one joint method. In Sections 3.2.1 and 3.2.2, 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 $\{\mathbf{\Omega}^{(g)}\}$ prior to the step of regression and then make use of the estimator to produce a better estimator of $\{\mathbf{B}^{(g)}\}$. In the plug-in method for inverse covariance estimation, we estimate $\{\mathbf{B}^{(g)}\}$ first and then incorporate the information to estimate $\{\mathbf{\Omega}^{(g)}\}$. In Section 3.2.3, we estimate $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$ together via double penalization.

3.2.1 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 $\{\hat{\mathbf{\Omega}}^{(g)}\}$ is available. Although $\mathbf{B}^{(g)}$ can be different for different g, we expect they share some common structure. 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, $\{\hat{\mathbf{\Omega}}^{(g)}\}$, available. Define $\boldsymbol{\beta}_{jk}$ as $\boldsymbol{\beta}_{jk} = (\beta_{jk}^{(1)}, ..., \beta_{jk}^{(G)})^T$. Regression parameters, $(\beta_{jk}^{(1)}, ..., \beta_{jk}^{(G)})$, 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

$$\underset{(\mathbf{B}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \operatorname{tr}\left\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\hat{\mathbf{\Omega}}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\} + \lambda_1 \sum_{j,k} p(\boldsymbol{\beta}_{jk}), \tag{3.3}$$

$$\text{where} \quad p(\boldsymbol{\beta}_{jk}) = \left(\sum_{g=1}^{G} \left|\beta_{jk}^{(g)}\right|\right)^{1/2}.$$

Here λ_1 is a tuning parameter. The penalty in (3.3) was proposed by Zhou and Zhu [2010],

which they call the hierarchical group penalty. This penalty controls the sparsity of $\{\hat{\mathbf{B}}^{(g)}\}$ 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 [2010] showed the penalty in (3.3) 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 (3.3). 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 $p(\boldsymbol{\beta}_{jk})$ for each group serves as a group penalty which encourages group shrinkage. Similar idea was previously considered by Turlach, Venables and Wright [2005], Yuan and Lin [2006], Zhang et al. [2008], and Zhao, Rocha and Yu [2009].

For the procedure in (3.3), we need to first estimate $\{\boldsymbol{\Omega}^{(g)}\}$. 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

$$\mathbf{S}^{(g)} = \frac{1}{n_g}(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g),0})(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g),0})^T. \tag{3.4}$$

Then $\{\hat{\boldsymbol{\Omega}}^{(g)}\}$ can be obtained by solving

$$\underset{\boldsymbol{\Omega}^{(g)}}{\operatorname{argmin}} \left\{ -\log\det(\boldsymbol{\Omega}^{(g)}) + \operatorname{tr}(\mathbf{S}^{(g)}\boldsymbol{\Omega}^{(g)}) + \lambda_2 \sum_{j\neq k} v_{jk}|\omega_{jk}| \right\}, \quad g = 1, ..., G. \tag{3.5}$$

The resulting solution is a sparse estimator of $\boldsymbol{\Omega}^{(g)}$. This technique was discussed in Section 1.2.

3.2.2 Plug-in Hierarchical Graphical LASSO estimator

In Section 3.2.1, we considered a plug-in method, PHL, which estimates $\{\boldsymbol{\Omega}^{(g)}\}$ first and then estimates $\{\mathbf{B}^{(g)}\}$ given $\{\hat{\boldsymbol{\Omega}}^{(g)}\}$. In this section, we propose the plug-in method using $\{\hat{\mathbf{B}}^{(g)}\}$ to estimate $\{\boldsymbol{\Omega}^{(g)}\}$. In particular, we first estimate $\{\mathbf{B}^{(g)}\}$ by using univariate regression techniques and obtain $\{\mathbf{S}^{(g)}\}$ defined in (3.4). With estimator $\{\mathbf{S}^{(g)}\}$ available, we propose a penalized likelihood method, the plug-in hierarchical graphical LASSO (PHGL) estimator,

by solving

$$\underset{(\mathbf{\Omega}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \left\{ -n_g \log\det(\mathbf{\Omega}^{(g)}) + n_g \operatorname{tr}(\mathbf{S}^{(g)}\mathbf{\Omega}^{(g)}) \right\} + \lambda_2 \sum_{s\neq t} \left(\sum_{g=1}^{G} \left| \omega_{st}^{(g)} \right| \right)^{1/2}, \tag{3.6}$$

where λ_2 is a tuning parameter $\mathbf{S}^{(g)} = \frac{1}{n_g}(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g)})(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g)})^T$.

This approach is closely related to the method previously considered by Guo et al. [2011]. 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 (3.6) is technically the same as that in their method, our resulting estimator has different graphical interpretations.

3.2.3 Doubly Penalized Sparse Estimator

In Sections 3.2.1 and 3.2.2, we considered two plug-in methods for estimation of $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$. In this section, we propose to estimate $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$ 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 $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$. We propose a joint penalized method, the doubly penalized sparse estimator (DPS), by solving

$$\underset{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \left\{ -l_g(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)}) + \lambda_1 \sum_{jk} \left(\sum_{g=1}^{G} \left| \beta_{jk}^{(g)} \right| \right)^{1/2} + \lambda_2 \sum_{s\neq t} \left(\sum_{g=1}^{G} \left| \omega_{st}^{(g)} \right| \right)^{1/2} \right\}, \tag{3.7}$$

where $l_g(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)}) = n_g \log\det(\mathbf{\Omega}^{(g)}) - \operatorname{tr}\left\{ (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{\Omega}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T \right\}$. As a group penalty, the first penalty term in (3.7) encourages the hierarchical sparsity among $\{\mathbf{B}^{(g)}\}$. In the meantime, the second penalty term in (3.7) serves as a group penalty for $\{\mathbf{\Omega}^{(g)}\}$.

Similar to the argument stated in Section 2.2.3, the objective function in (3.7) is not convex with respect to $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$ and the optimization can be unstable when $\max\{n_1, ..., n_G\} < p$. With diagonal $\{\mathbf{\Omega}^{(g)}\}$, the first term in $l_g(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})$ can dominate the other terms in the objective function if the trace terms are zero, which may occur when $\max\{n_1, ..., n_G\} < p$. Therefore, the objective function can keep decreasing as the diagonal entries in $\{\mathbf{\Omega}^{(g)}\}$ continue to increase. Therefore, if $\max\{n_1, ..., n_G\} < p$, the plug-in methods in Sections 3.2.1 and 3.2.2 are recommended and can often perform better than the DPS

method.

3.2.4 Model Selection

In Sections 3.2.1 - 3.2.3, we proposed two plug-in methods and one joint method for estimation of $\{(\mathbf{B}^{(g)}, \mathbf{\Omega}^{(g)})\}$. To apply these methods, we first need to select the tuning parameters λ_1 and λ_2 in (3.3), (3.6), and (3.7), which control the sparsity of the resulting estimators. The tuning parameters can be selected either using validation sets or through K-fold cross-validation as stated in Section 2.2.4. In particular, 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 $\{(\hat{\mathbf{B}}_{\lambda_1,(-k)}^{(g)}, \hat{\mathbf{\Omega}}_{\lambda_2,(-k)}^{(g)})\}$. For the PHL method, the optimal tuning parameter $\hat{\lambda}_1$ is selected which minimizes the prediction error defined by

$$\mathrm{CV}(\lambda_1) = \sum_{k=1}^{K} \sum_{g=1}^{G} \| \mathbf{Y}_{(k)}^{(g)} - \mathbf{X}_{(k)}^{(g)} \hat{\mathbf{B}}_{\lambda_1,(-k)}^{(g)} \|_F^2, \tag{3.8}$$

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

$$\mathrm{CV}(\lambda_2) = \sum_{k=1}^{K} \sum_{g=1}^{G} \Big[n_{(g,k)} \log\det(\hat{\mathbf{\Omega}}_{\lambda_2,(-k)}^{(g)}) - \mathrm{tr}\left\{ (\mathbf{Y}_{(k)}^{(g)} - \mathbf{X}_{(k)}^{(g)} \hat{\mathbf{B}}^{(g)}) \hat{\mathbf{\Omega}}_{\lambda_2,(-k)}^{(g)} (\mathbf{Y}_{(k)}^{(g)} - \mathbf{X}_{(k)}^{(g)} \hat{\mathbf{B}}^{(g)})^T \right\} \Big], \tag{3.9}$$

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 (3.8) with a prespecified λ_2 and the optimal $\hat{\lambda}_2$ is selected using (3.9) 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 $\hat{\lambda}_1$.

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 $\{(\mathbf{X}_{(k)}^{(g)}, \mathbf{Y}_{(k)}^{(g)})\}$ in (3.8) and (3.9), 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.

3.3 Asymptotic Properties

In this section, we investigate the asymptotic behavior of our three proposed methods in Sections 3.2.1 - 3.2.3 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 [2001], Yuan and Lin [2007] and Zou [2006]. The technical derivation uses the results in Knight and Fu [2000]. 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 $A = [a_1, ..., a_p]$ by $\text{Vec}(A) = (a_1^T, ..., a_p^T)^T$. Let $\boldsymbol{\beta}^* = (\text{Vec}(\mathbf{B}^{*,(1)})^T, ..., \text{Vec}(\mathbf{B}^{*,(G)})^T)^T$ be the true regression parameter vector and $\boldsymbol{\omega}^* = (\text{Vec}(\boldsymbol{\Omega}^{*,(1)})^T, ..., \text{Vec}(\boldsymbol{\Omega}^{*,(G)})^T)^T$ 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 (3.3).

Theorem 4. *Suppose that $\lambda_1 n^{-\frac{1}{2}} \to 0$ as $n \to \infty$ and $\hat{\boldsymbol{\Omega}}^{(g)}$ in (3.3) is a consistent estimator of $\boldsymbol{\Omega}^{*,(g)}; g = 1, ..., G$. 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 (3.3) such that $\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \| = O_p(\frac{1}{\sqrt{n}})$, where $\hat{\boldsymbol{\beta}} = (\,Vec(\hat{\mathbf{B}}^{(1)})^T, ..., Vec(\hat{\mathbf{B}}^{(G)})^T)^T$;*

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

Theorem 4 states that with a consistent estimator of $\boldsymbol{\Omega}^{*,(g)}$, 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 5. *Suppose that $\lambda_2 n^{-\frac{1}{2}} \to 0$ as $n \to \infty$ and $\hat{\mathbf{B}}^{(g)}$ in (3.6) is a $\sqrt{n}$-consistent estimator of $\mathbf{B}^{*,(g)}; g = 1, ..., G$.*

1. *(Consistency) There exists a local minimizer of* (3.6) *such that* $\| \hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^* \| = O_p(\frac{1}{\sqrt{n}})$, *where* $\hat{\boldsymbol{\omega}} = (\, Vec(\hat{\boldsymbol{\Omega}}^{(1)})^T, ..., Vec(\hat{\boldsymbol{\Omega}}^{(G)})^T)^T$;

2. *(Sparsity) If* $\lambda_2 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{\omega}_{jk}^{(g)} = 0) = 1$ *if* $\omega_{jk}^{*,(g)} = 0$.

In theorems 4 and 5, we establish the consistency and sparsity of plug-in estimators. The following theorem shows the similar asymptotic properties of the DPS solution in which $\{\hat{\mathbf{B}}^{(g)}\}$ and $\{\hat{\boldsymbol{\Omega}}^{(g)}\}$ are obtained together.

Theorem 6. *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;* $g = 1, ..., G$.

1. *(Consistency) There exists a local minimizer of* (3.7) *such that*
$\| (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\omega}}^T)^T - (\boldsymbol{\beta}^{*T}, \boldsymbol{\omega}^{*T})^T \| = O_p(\frac{1}{\sqrt{n}})$,
where $\hat{\boldsymbol{\beta}} = (\, Vec(\hat{\mathbf{B}}^{(1)})^T, ..., Vec(\hat{\mathbf{B}}^{(G)})^T)^T$ *and* $\hat{\boldsymbol{\omega}} = (\, Vec(\hat{\boldsymbol{\Omega}}^{(1)})^T, ..., Vec(\hat{\boldsymbol{\Omega}}^{(G)})^T)^T$;

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

3. *(Sparsity of* $\{\hat{\boldsymbol{\Omega}}^{(g)}\}$ *) If* $\lambda_2 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{\omega}_{jk}^{(g)} = 0) = 1$ *if* $\omega_{jk}^{*,(g)} = 0$.

3.4 Computational Algorithm

In this section, we describe computational algorithms to solve problems (3.3), (3.6), and (3.7). In particular, we apply the coordinate-descent algorithms described in Section 2.4 iteratively, with combination of the local linear approximation (LLA) [Zou and Li, 2008].

We now describe the algorithm for the PHL method in details. Denote the estimates of $\beta_{jk}^{(g)}$ from the i-th iteration by $(\hat{\beta}_{jk}^{(g)})^{(i)}$. Then by applying the LLA, the penalty term in (3.3) at the $(i+1)$-th iteration can be approximated as follows,

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

Then, at the $(i+1)$-th iteration, the problem (3.3) is decomposed into G individual optimization problems

$$\underset{\mathbf{B}^{(g)}}{\operatorname{argmin}} \operatorname{tr} \left\{ (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)}) \hat{\boldsymbol{\Omega}}^{(g)} (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T \right\} + \lambda_1 \sum_{j,k} w_{jk} \left| \beta_{jk}^{(g)} \right|, \tag{3.10}$$

where $w_{jk} = \frac{1}{2}\left(\sum_{g=1}^{G}\left|(\hat{\beta}_{jk}^{(g)})^{(i)}\right|\right)^{-1/2}$ and $g = 1, ..., G$. The optimization problem (3.10) 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 in Section 2.2.1. Therefore, the algorithm for (3.3) proceeds as follows:

Algorithm for the PHL Method

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

Step 2 (Updating rule). For $g = 1, ..., G$, update $\{(\hat{\mathbf{B}}^{(g)})^{(i)}\}$ by applying the coordinate-descent algorithm for the the plug-in weighted LASSO method in Section 2.2.1 to the problem (3.10).

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

Next we describe the algorithm for the PHGL method in Section 3.2.2. Similar to the algorithm for the PHL method, we first apply the LLA to the objective function in (3.6) with the current estimates $\{(\hat{\mathbf{\Omega}}^{(g)})^{(i)}\}$. Then, at the $(i+1)$-th iteration, the problem (3.6) is decomposed into G individual optimization problems

$$\underset{\mathbf{\Omega}^{(g)}}{\operatorname{argmin}}\left\{-n_g \log\det(\mathbf{\Omega}^{(g)}) + n_g \operatorname{tr}(\mathbf{S}^{(g)}\mathbf{\Omega}^{(g)})\right\} + \lambda_2 \sum_{s\neq t} v_{st}\left|\omega_{st}^{(g)}\right|, \tag{3.11}$$

where $v_{st} = \frac{1}{2}\left(\sum_{g=1}^{G}\left|(\hat{\omega}_{jk}^{(g)})^{(i)}\right|\right)^{-1/2}$ and $g = 1, ..., G$. The problem (3.11) can be solved by applying the GLASSO algorithm. Therefore, the algorithm for (3.6) proceeds as follows:

Algorithm for the PHGL Method

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

Step 2 (Updating rule). For $g = 1, ..., G$, update $\{(\hat{\mathbf{\Omega}}^{(g)})^{(i)}\}$ by applying the GLASSO algorithm to the problem (3.11).

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

Next, we combine the above two algorithms to solve problem (3.7) 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 $\{\mathbf{\Omega}^{(g)}\}$). Set the separate LASSO solution $\{(\hat{\mathbf{B}}^{(g)})^{(i)}\}$

as the initial value for $\{\mathbf{B}^{(g)}\}$ and the separate GLASSO solution $\{(\hat{\mathbf{\Omega}}^{(g)})^{(i)}\}$ as the initial value of $\{\mathbf{\Omega}^{(g)}\}$.

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

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

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

As we point out in Section 3.2.3, when $\max\{n_1, ..., n_G\} < p$, the solution can possibly be unstable with very small residual variances. In that case, the plug-in methods may perform better.

3.5 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 in Section 2.2.3 separately to each group. The estimator is given by solving (3.2). 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 $\{\mathbf{\Omega}^{(g)}\}$. In particular, the estimator of $\{\mathbf{\Omega}^{(g)}\}$ is given by solving (1.6). The resulting estimator of $\{\mathbf{B}^{(g)}\}$ will be called the LASSO estimator and the resulting estimator of $\{\mathbf{\Omega}^{(g)}\}$ 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 $\{\mathbf{\Omega}^{(g)}\}$ jointly. Once we have the estimator of $\{\mathbf{\Omega}^{(g)}\}$, 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 $\{\mathbf{\Omega}^{(g)}\}$ 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 $\mathbf{\Omega}$ common in all groups. Figure 3.3 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 $\mathbf{\Omega}^{(g)}$, 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 $\mathbf{N}(\mathbf{B}^{(g)^T}\mathbf{x}_i^{(g)}, \mathbf{\Omega}^{(g)})$, where $\mathbf{x}_i^{(g)}; i = 1, ..., n$ are i.i.d vectors from $\mathbf{N}(0, I_p)$.

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

$$\text{PE} = \frac{1}{nmG} \sum_{g=1}^{G} \| \mathbf{Y}^{(g)} - \hat{\mathbf{Y}}^{(g)} \|_F^2,$$

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

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

$$\text{EL} = \frac{1}{G} \sum_{g=1}^{G} \left[\text{tr}(\mathbf{\Sigma}^{(g)} \hat{\mathbf{\Omega}}^{(g)}) - \log(|\mathbf{\Sigma}^{(g)} \hat{\mathbf{\Omega}}^{(g)}|) - m \right],$$

$$\text{FL} = \frac{1}{G} \sum_{g=1}^{G} \| \mathbf{\Omega}^{(g)} - \hat{\mathbf{\Omega}}^{(g)} \|_F^2 / \| \mathbf{\Omega}^{(g)} \|_F^2.$$

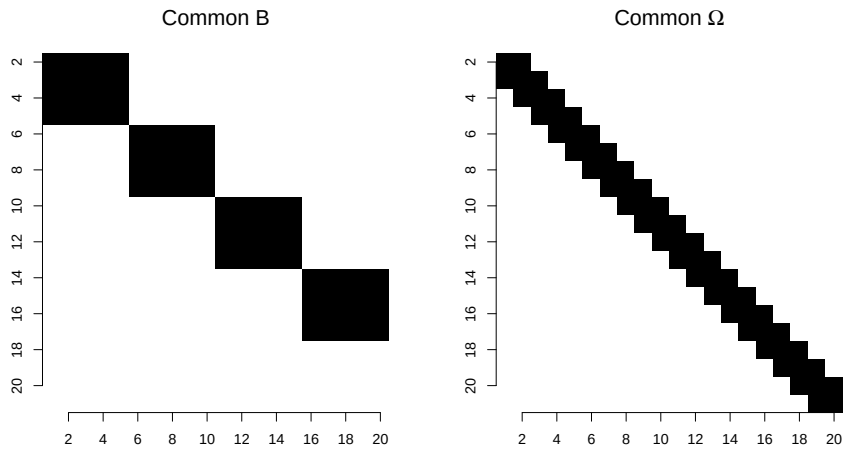

Figure 3.3: Regression Parameter Structure and Inverse Covariance Structure that are common in all groups. Non-zero entries are colored as black and zero entries are colored as white.

Table 3.1: 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 3.1 and Figures 3.4-3.6 summarize the results. When $\rho = 0$, M2 outperforms the others in both prediction and estimation of $\{\mathbf{\Omega}^{(g)}\}$. 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 $\{\mathbf{\Omega}^{(g)}\}$ when all groups share some common structure.

Table 3.2 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

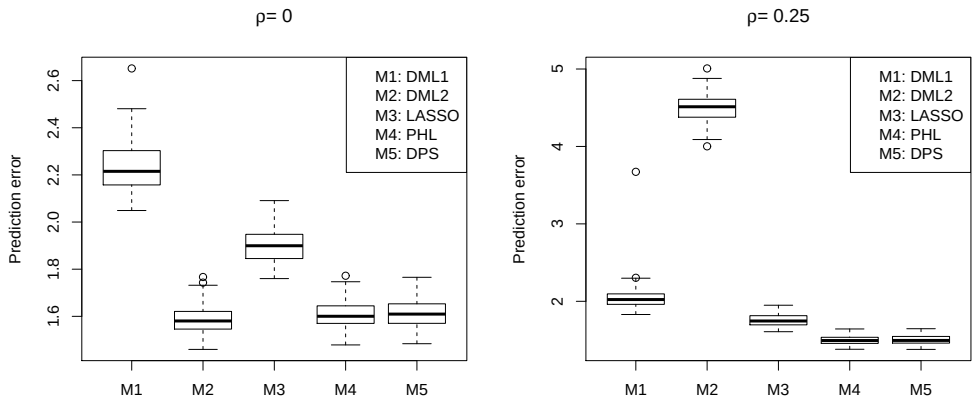

Figure 3.4: Boxplots of prediction errors of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

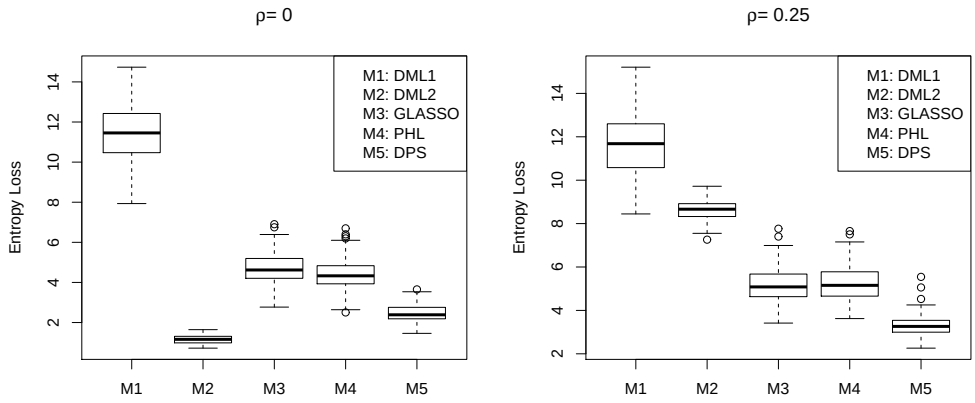

Figure 3.5: Boxplots of entropy losses of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

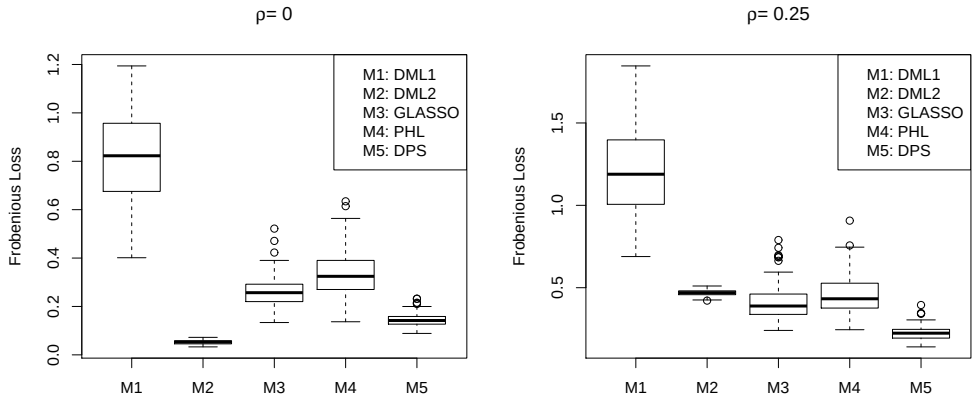

Figure 3.6: Boxplots of Frobenius losses of all methods based on 100 replications. Left: All groups are the same. Right: There exist the common and unique structures across groups.

Table 3.2: 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)

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 $\{\mathbf{\Omega}^{(g)}\}$ in our simulated examples.

3.6 Application to the Glioblastoma Cancer Data

In this section, we apply our proposed methods to the GBM cancer dataset. In this dataset, there are 17814 genes and 534 microRNAs 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 [Verhaak et al., 2010]. One important goal is to regress genes on microRNAs to investigate the effect of microRNAs on gene expressions. The other goal is to estimate the conditional inverse covariance matrix of gene expressions given microRNAs. This matrix can help us to interpret the conditional relationship among genes given microRNAs.

To proceed with the analysis, preprocessing is necessary. There are many possibilities for preprocessing. For example, Bair and Tibshirani [2004] 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. [2010] established 840 signature genes which are highly distinctive for four subtypes. The expression levels of these genes are depicted in Figure 3.7. We use these 840 signature genes to explore distinctive effects of microRNAs 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

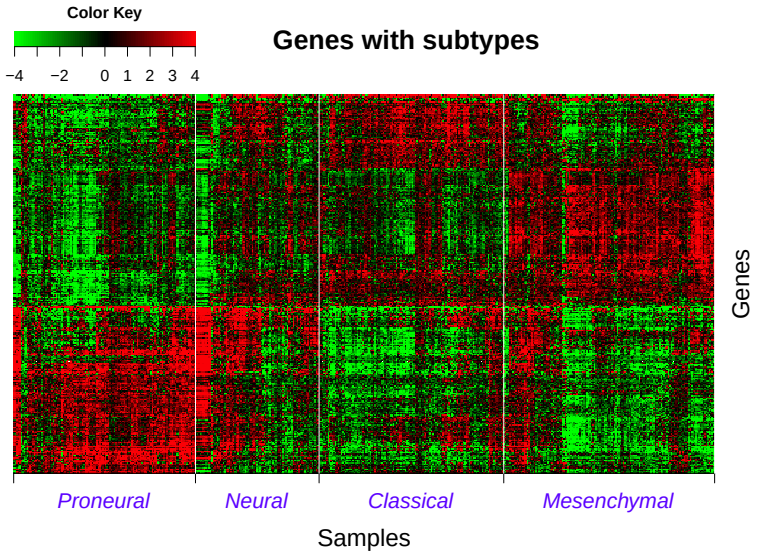

Figure 3.7: Heatmap of expression levels of 840 signature genes established by Verhaak et al. [2010].

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 [2005]. WGCNA detects modules using a hierarchical clustering method with the topological overlap dissimilarity measure [Ravasz et al., 2002]. Zhang and Horvath [2005] 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. [2010] 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 microRNAs which are predicted to target at least one of the selected genes and have large MADs. As a result, 40 genes and 50 microRNAs are used for the results in this analysis.

Table 3.3: 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)

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 in Section 2.2.3 to the combined data. In particular, the estimator can be obtained by solving (3.2). In the second approach, we apply LASSO and GLASSO. The detailed description of this approach is presented in M3 in Section 3.5. 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 3.2.4. We perform the random splitting 100 times. By using the test set, we assess prediction performance of several methods including our proposed methods.

Table 3.3 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 3.8 shows the averaged estimated regression coefficients over 100 replications of several microRNAs for some selected genes. In order to produce the heatmap, the DPS estimates are used. The results show some interesting relationships between genes and microRNAs that are specific to certain GBM subtypes. For instance, we have observed a

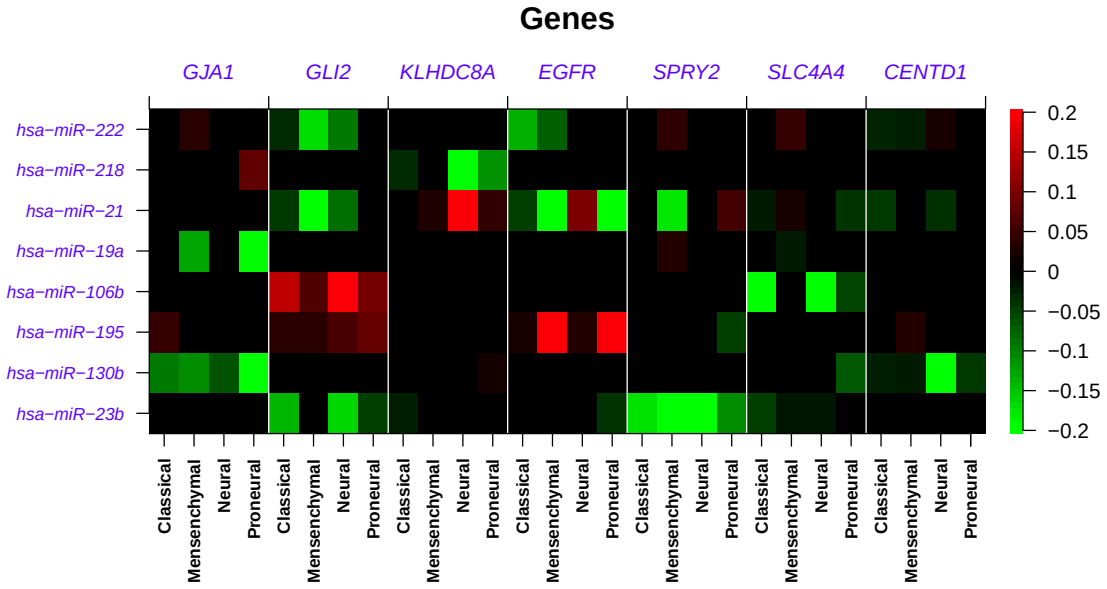

Figure 3.8: Heatmap of averaged estimated regression coefficients of several microRNAs for some selected genes. The DPS estimates are used to generate the heatmap.

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 [Elsawa et al., 2011]. It has been shown that the knockdown of *GLI2* mRNA has significantly decreased the migratory ability of human glioblastoma cells [Uchida et al., 2011]. Herein, our results suggest that the accelerated inflammatory response observed in the GBM Mesenchymal subtype might be partially through miR222-dependent GLI2 regulation [Verhaak et al., 2010].

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. [2006] 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 microRNAs 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

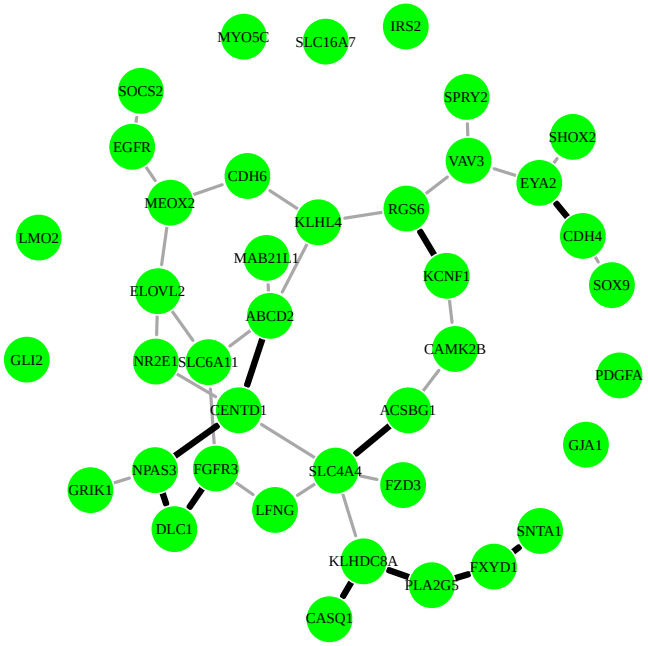

Figure 3.9: A graphical model of gene expressions based on the estimated inverse covariance matrix. Black lines are common edges across all subgroups. Grey lines are unique edges to some subgroups. The DPS estimates are used to generate the network.

papers have shown that *EGFR* regulates *miR21* in a couple of cancers, including human glioblastoma and lung cancers [Zhou et al., 2010; Seike et al., 2009]. 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 3.9 shows the estimated conditional inverse covariance structure of genes given microRNAs. 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. [2010] 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 3.9 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 microRNAs 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 microRNAs

among different cancer subtypes.

3.7 Discussion

In this chapter, 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 3.2.3. 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. 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, Lafferty and Wasserman [2009] and Cai, Liu and Luo [2011]. Another research direction is to extend our methods to non-Gaussian situations. Further exploration is needed.

3.8 Proofs

3.8.1 Proof of Theorem 4

Consistency

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

$$Q(\boldsymbol{\beta}) = \sum_{g=1}^{G} \text{tr}\left\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\hat{\boldsymbol{\Omega}}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\} + \lambda_1 \sum_{j,k} p(\boldsymbol{\beta}_{jk}). \tag{3.12}$$

To show the results, we use the similar technique in the proof of Theorem 1 in Fan and Li [2001]. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\left\{\sup_{\|U\|=D} Q(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}U) > Q(\boldsymbol{\beta}^*)\right\} > 1 - \delta, \tag{3.13}$$

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

Let $\mathbf{y}^{(g)} = \text{Vec}(\mathbf{Y}^{(g)})$, $\mathbf{X}^{m,(g)} = \mathbf{I}_m \otimes \mathbf{X}^{(g)}$ and $\boldsymbol{\beta}^{(g)} = \text{Vec}(\mathbf{B}^{(g)}); g = 1, ..., G$. Then we can rewrite $Q(\boldsymbol{\beta})$ in (3.12) as

$$Q(\boldsymbol{\beta}) = \sum_{g=1}^{G} (\mathbf{y}^{(g)} - \mathbf{X}^{m,(g)}\boldsymbol{\beta}^{(g)})^T(\hat{\boldsymbol{\Omega}}^{(g)} \otimes \mathbf{I}_n)(\mathbf{y}^{(g)} - \mathbf{X}^{m,(g)}\boldsymbol{\beta}^{(g)}) + \lambda_1 \sum_{j,k} p(\boldsymbol{\beta}_{jk}). \tag{3.14}$$

Define $V_n(U) = Q(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}U) - Q(\boldsymbol{\beta}^*)$. Using (3.14), we can show that

$$\begin{aligned} V_n(U) = &\sum_{g=1}^{G} U^{(g)T}(\hat{\boldsymbol{\Omega}}^{(g)} \otimes \frac{1}{n}\mathbf{X}^{(g)T}\mathbf{X}^{(g)})U^{(g)} - \sum_{g=1}^{G} 2\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\boldsymbol{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)} \\ &+ \lambda_1 \sum_{j,k}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} - \left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}, \end{aligned} \tag{3.15}$$

where $\boldsymbol{\epsilon}^{(g)} = \text{Vec}(\mathbf{e}^{(g)}); g = 1, ..., G$. Define $\mathcal{I} = \{(j,k) | \beta_{jk}^{*,(g)} \neq 0 \text{ for some } g = 1, ..., G\}$. Since $\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} - \left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2} = \left(\sum_{g=1}^{G}\left|\frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} \geq 0$ for $(j,k) \notin \mathcal{I}$, we have that

$$\begin{aligned} V_n(U) \geq &\sum_{g=1}^{G} U^{(g)T}(\hat{\boldsymbol{\Omega}}^{(g)} \otimes \frac{1}{n}\mathbf{X}^{(g)T}\mathbf{X}^{(g)})U^{(g)} - \sum_{g=1}^{G} 2\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\boldsymbol{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)} \\ &+ \lambda_1 \sum_{(j,k)\in\mathcal{I}}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} - \left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}. \end{aligned} \tag{3.16}$$

For the first term on the right-hand side of (3.16), note that

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

as $\frac{1}{n}\mathbf{X}^{(g)T}\mathbf{X}^{(g)} \to A^{(g)}$ and $\hat{\mathbf{\Omega}}^{(g)} \to_p \mathbf{\Omega}^{*,(g)}; g = 1, ..., G$.

For the second term on the right-hand side of (3.16), note that

$$\begin{aligned}\left|\sum_{g=1}^{G} 2\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\mathbf{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)}U^{(g)}\right| &\leq 2\sum_{g=1}^{G} \parallel \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\mathbf{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)} \parallel \parallel U^{(g)} \parallel \\ &\leq 2\sum_{g=1}^{G} \parallel \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\mathbf{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)} \parallel \parallel U \parallel \\ &= O_p(1) \parallel U \parallel\end{aligned}$$

as $\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^{(g)T}(\hat{\mathbf{\Omega}}^{(g)} \otimes \mathbf{I}_n)\mathbf{X}^{m,(g)} \to_d Z$ where Z has multivariate normal distribution.

For the third term on the right-hand side of (3.16), we can show that

$$\begin{aligned}&\lambda_1 \sum_{(j,k)\in\mathcal{I}} \left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} - \left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\} \\ &= \lambda_1 \sum_{(j,k)\in\mathcal{I}} \sum_{g=1}^{G} \frac{1}{\gamma_{jk}}\left\{\left|\beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right| - \left|\beta_{jk}^{*,(g)}\right|\right\} \\ &= \frac{\lambda_1}{\sqrt{n}} \sum_{(j,k)\in\mathcal{I}} \sum_{g=1}^{G} \frac{1}{\gamma_{jk}}\left\{\left|u_{jk}^{(g)}\right|\text{sign}(\beta_{jk}^{*,(g)}) + o(1)\right\} = o_p(1),\end{aligned}$$

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

By combining above statements, we have

$$V_n(U) \geq \sum_{g=1}^{G} U^{(g)T}(\mathbf{\Omega}^{*,(g)} \otimes A^{(g)})U^{(g)} + O_p(1) \parallel U \parallel + o_p(1).$$

By choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \parallel U \parallel = D\}$ with the probability greater than $1 - \delta$ as $\mathbf{\Omega}^{*,(g)}$ and $A^{(g)}$ are positive-definite. Therefore, (3.13) holds. This completes the proof of the consistency.

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 (3.12) with respect to $\beta_{jk}^{(g)}$ at $\hat{\beta}_{jk}^{(g)}$ has the same sign as $\hat{\beta}_{jk}^{(g)}$. Let $\boldsymbol{\beta}^{(g)} = \text{Vec}(\mathbf{B}^{(g)})$ and note that

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}^{(\mathrm{g})}}(\mathbf{y}^{(\mathrm{g})} - \mathbf{X}^{m,(\mathrm{g})}\boldsymbol{\beta}^{(\mathrm{g})})^T(\hat{\boldsymbol{\Omega}}^{(\mathrm{g})} \otimes \mathbf{I}_n)(\mathbf{y}^{(\mathrm{g})} - \mathbf{X}^{m,(\mathrm{g})}\boldsymbol{\beta}^{(\mathrm{g})})\big|_{\boldsymbol{\beta}^{(\mathrm{g})}=\hat{\boldsymbol{\beta}}^{(\mathrm{g})}} \\
&= (\hat{\boldsymbol{\Omega}}^{(\mathrm{g})} \otimes \mathbf{X}^{(\mathrm{g})^T}\mathbf{X}^{(\mathrm{g})})(\hat{\boldsymbol{\beta}}^{(\mathrm{g})} - \boldsymbol{\beta}^{*,(g)}) - (\hat{\boldsymbol{\Omega}}^{(\mathrm{g})} \otimes \mathbf{X}^{(\mathrm{g})^T})\boldsymbol{\epsilon}^{(\mathrm{g})} \\
&= \sqrt{n}\left\{(\hat{\boldsymbol{\Omega}}^{(\mathrm{g})} \otimes \frac{1}{n}\mathbf{X}^{(\mathrm{g})^T}\mathbf{X}^{(\mathrm{g})})\sqrt{n}(\hat{\boldsymbol{\beta}}^{(\mathrm{g})} - \boldsymbol{\beta}^{*,(g)}) - \frac{1}{\sqrt{n}}(\hat{\boldsymbol{\Omega}}^{(\mathrm{g})} \otimes \mathbf{X}^{(\mathrm{g})^T})\boldsymbol{\epsilon}^{(\mathrm{g})}\right\} \\
&= \sqrt{n}O_p(1)
\end{aligned}
$$

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

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

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

3.8.2 Proof of Theorem 5

Consistency

Let $\boldsymbol{\omega} = (\mathrm{Vec}(\boldsymbol{\Omega}^{(1)})^T, ..., \mathrm{Vec}(\boldsymbol{\Omega}^{(G)})^T)^T$ and define $Q(\boldsymbol{\omega})$ as

$$
Q(\boldsymbol{\omega}) = \sum_{g=1}^{G}\left\{-n\log\det(\boldsymbol{\Omega}^{(\mathrm{g})}) + n\mathrm{tr}(\mathbf{S}^{(\mathrm{g})}\boldsymbol{\Omega}^{(\mathrm{g})})\right\} + \lambda_2\sum_{s\neq t}\left(\sum_{g=1}^{G}\left|\omega_{st}^{(\mathrm{g})}\right|\right)^{1/2} \tag{3.17}
$$

To show the results, we use the similar technique in the proof of Theorem 4. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$
P\left\{\sup_{\|U\|=D} Q(\boldsymbol{\omega}^* + \frac{1}{\sqrt{n}}U) > Q(\boldsymbol{\omega}^*)\right\} > 1 - \delta, \tag{3.18}
$$

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

Using (3.17), define $V_n(U)$ as

$$
\begin{aligned}
V_n(U) &= Q(\boldsymbol{\omega}^* + \frac{1}{\sqrt{n}}U) - Q(\boldsymbol{\omega}^*) \\
&= \sum_{g=1}^{G}\left\{-n\log\det((\boldsymbol{\Omega}^{*,(\mathrm{g})} + \frac{U^{(\mathrm{g})}}{\sqrt{n}})(\boldsymbol{\Omega}^{*,(\mathrm{g})})^{-1}) + n\mathrm{tr}(\frac{U^{(\mathrm{g})}\mathbf{S}^{(\mathrm{g})}}{\sqrt{n}})\right\} \\
&+ \lambda_2\sum_{s\neq t}\left\{\left(\sum_{g=1}^{G}\left|\omega_{st}^{*,(g)} + \frac{1}{\sqrt{n}}u_{jk}^{(g)}\right|\right)^{1/2} - \left(\sum_{g=1}^{G}\left|\omega_{jk}^{*,(g)}\right|\right)^{1/2}\right\}.
\end{aligned}
$$

Using the similar argument as in the proof of Lemma 2, it can be shown that

$$V_n(U) = \sum_{g=1}^{G} \mathrm{tr}(U^{(g)}\mathbf{\Sigma}^{(g)}U^{(g)}\mathbf{\Sigma}^{(g)}) + \sum_{g=1}^{G} \mathrm{tr}[U^{(g)}\sqrt{n}(\mathbf{S}^{(g)} - \mathbf{\Sigma}^{(g)})]$$
$$+ \lambda_2 \sum_{s\neq t} \left\{ \left(\sum_{g=1}^{G} \left| \omega_{st}^{*,(g)} + \frac{1}{\sqrt{n}} u_{jk}^{(g)} \right| \right)^{1/2} - \left(\sum_{g=1}^{G} \left| \omega_{jk}^{*,(g)} \right| \right)^{1/2} \right\} + o(1). \tag{3.19}$$

For the second term on the right-hand side of (3.19), by using the similar argument as in the proof of Theorem 2, it can be shown that

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

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

For the third term on the right-hand side of (3.19), by using the similar argument as in the proof of Theorem 4, it can be shown that

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

By combining the above statements, we can conclude that the first term on the right-hand side of (3.19) 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.

Sparsity

Similar to the proof of the sparsity in Theorem 4, it is sufficient to show that with probability tending to 1 as $n \to \infty$, for any (s,t) such that $\omega_{st}^{*,(g)} = 0$, the partial derivative of Q in (3.17) with respect to $\omega_{st}^{(g)}$ at $\hat{\omega}_{st}^{(g)}$ has the same sign as $\hat{\omega}_{st}^{(g)}$. Note that

$$\frac{\partial Q}{\partial \omega_{st}^{(g)}}\Big|_{\omega_{st}^{(g)}=\hat{\omega}_{st}^{(g)}} = n(\boldsymbol{s}_{st}^{(g)} - \hat{\sigma}_{st}^{(g)}) + \lambda_2 \frac{\mathrm{sign}(\hat{\omega}_{st}^{(g)})}{2(\sum_{g=1}^{G} |\hat{\omega}_{st}^{(g)}|)^{1/2}}$$

where $\mathbf{S}^{(g)} = (\boldsymbol{s}_{st}^{(g)})$ and $(\hat{\mathbf{\Omega}}^{(g)})^{-1} = (\hat{\sigma}_{st}^{(g)})$. By using the argument in the proof of Theorem 2 in Guo et al. [2011], one can show that $(\boldsymbol{s}_{st}^{(g)} - \hat{\sigma}_{st}^{(g)}) = O_p(1/\sqrt{n})$. Therefore, we have

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

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

3.8.3 Proof of Theorem 6

Consistency

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

$$Q(\boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{g=1}^{G} \left\{ -l_g(\boldsymbol{\beta}, \boldsymbol{\omega}) + \lambda_1 \sum_{jk} \left(\sum_{g=1}^{G} \left| \beta_{jk}^{(g)} \right| \right)^{1/2} + \lambda_2 \sum_{s \neq t} \left(\sum_{g=1}^{G} \left| \omega_{st}^{(g)} \right| \right)^{1/2} \right\}, \tag{3.20}$$

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

To show the results, we use the similar technique in the proof of Theorem 4. It suffices to show that for any given $\delta > 0$, there exists a large constant D such that

$$P\left\{ \sup_{\|U\|=D} Q(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}} U_1, \boldsymbol{\omega}^* + \frac{1}{\sqrt{n}} U_2) > Q(\boldsymbol{\beta}^*, \boldsymbol{\omega}^*) \right\} > 1 - \delta, \tag{3.21}$$

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

Using (3.20), define $V_n(U) = Q(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}} U_1, \boldsymbol{\omega}^* + \frac{1}{\sqrt{n}} U_2) - Q(\boldsymbol{\beta}^*, \boldsymbol{\omega}^*)$. It can be shown that

$$\begin{aligned}
V_n(U) = & \sum_{g=1}^{G} \left\{ -n \log\det((\boldsymbol{\Omega}^{*,(g)} + \frac{U_2^{(g)}}{\sqrt{n}})(\boldsymbol{\Omega}^{*,(g)})^{-1}) + n\mathrm{tr}(\frac{U_2^{(g)} \mathbf{S}^{*,(g)}}{\sqrt{n}}) \right\} \\
& + \sum_{g=1}^{G} \left\{ \mathrm{tr}\left[(\boldsymbol{\Omega}^{*,(g)} + \frac{U_2^{(g)}}{\sqrt{n}}) (\frac{\mathbf{X}^{(g)} U_1^{(g)}}{\sqrt{n}})^T (\frac{\mathbf{X}^{(g)} U_1^{(g)}}{\sqrt{n}}) \right] \right\} \\
& - 2 \sum_{g=1}^{G} \left\{ \mathrm{tr}\left[(\boldsymbol{\Omega}^{*,(g)} + \frac{U_2^{(g)}}{\sqrt{n}}) (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{*,(g)})^T (\frac{\mathbf{X}^{(g)} U_1^{(g)}}{\sqrt{n}}) \right] \right\} \\
& + \lambda_1 \sum_{j,k} \left\{ \left(\sum_{g=1}^{G} \left| \beta_{jk}^{*,(g)} + \frac{1}{\sqrt{n}} u_{1,jk}^{(g)} \right| \right)^{1/2} - \left(\sum_{g=1}^{G} \left| \beta_{jk}^{*,(g)} \right| \right)^{1/2} \right\} \\
& + \lambda_2 \sum_{s \neq t} \left\{ \left(\sum_{g=1}^{G} \left| \omega_{st}^{*,(g)} + \frac{1}{\sqrt{n}} u_{2,jk}^{(g)} \right| \right)^{1/2} - \left(\sum_{g=1}^{G} \left| \omega_{jk}^{*,(g)} \right| \right)^{1/2} \right\}.
\end{aligned} \tag{3.22}$$

For the first term on the right-hand side of (3.22), it has been shown in Theorem 4 that

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

For the second term and the third term on the right-hand side of (3.22), by using the

similar argument in the proof of Lemma 3, it can be shown that

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

and

$$\sum_{g=1}^{G}\left\{\mathrm{tr}\left[(\mathbf{\Omega}^{*,(g)}+\frac{U_2^{(g)}}{\sqrt{n}})(\mathbf{Y}^{(g)}-\mathbf{X}^{(g)}\mathbf{B}^{*,(g)})^T(\frac{\mathbf{X}^{(g)}U_1^{(g)}}{\sqrt{n}})\right]\right\}=O_p(1).$$

For the fourth and fifth term on the right-hand side of (3.22), it has been shown in Theorems 4 and 5 that

$$\lambda_1\sum_{j,k}\left\{\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}+\frac{1}{\sqrt{n}}u_{1,jk}^{(g)}\right|\right)^{1/2}-\left(\sum_{g=1}^{G}\left|\beta_{jk}^{*,(g)}\right|\right)^{1/2}\right\}=o_p(1),$$

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

By combining the above statements, we can conclude that the right-hand side of (3.22) is dominated by $\sum_{g=1}^{G}\mathrm{tr}(U_2^{(g)}\mathbf{\Sigma}^{(g)}U_2^{(g)}\mathbf{\Sigma}^{(g)})$ and $\sum_{g=1}^{G}{U_1^{(g)}}^T(\mathbf{\Omega}^{*,(g)}\otimes A^{(g)})U_1^{(g)}$. Therefore, by choosing a sufficiently large D, $V_n(U)>0$ uniformly on $\{U:\parallel U\parallel=D\}$ with the probability greater than $1-\delta$. This completes the proof of the consistency.

Sparsity

Note that $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\omega}})$ is a $\sqrt{n}$-consistent local minimizer of $Q(\boldsymbol{\beta},\boldsymbol{\omega})$ defined in (3.20). As $\hat{\boldsymbol{\beta}}=\mathrm{argmin}_{\boldsymbol{\beta}}Q(\boldsymbol{\beta},\hat{\boldsymbol{\omega}})$ and $\hat{\boldsymbol{\omega}}$ is $\sqrt{n}$-consistent, the sparsity of $\hat{\boldsymbol{\beta}}$ holds by Theorem 4. Similarly, since $\hat{\boldsymbol{\omega}}=\mathrm{argmin}_{\boldsymbol{\omega}}Q(\hat{\boldsymbol{\beta}},\boldsymbol{\omega})$ and $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, the sparsity of $\hat{\boldsymbol{\omega}}$ holds by Theorem 5. These complete the proof of this theorem.

Chapter 4

Joint Estimation of Multiple Precision Matrices

4.1 Introduction

In Chapters 2-3, we considered simultaneous modeling of the regression coefficient matrix and the inverse of residual covariance matrix. In this chapter, we mainly focus on the estimation of precision matrices.

As stated in Section 1.2, estimation of a precision matrix, which is an inverse covariance matrix, has attracted a lot of attention recently in the context of the Gaussian graphical model. The precision matrix also plays an important role in other various areas of statistical analysis. For example, some classification techniques such as linear discriminant analysis and quadratic discriminant analysis require good estimates of precision matrices.

All approaches in Section 1.2 focus on estimation of a single precision matrix. The fundamental assumption of these approaches is that all observations follow the same distribution. However, in some real applications, this assumption can be unreasonable. For instance, as pointed out in Section 3.1, the GBM cancer can be classified into four subtypes [Verhaak et al., 2010]. In this case, it would be more realistic to assume that the distribution of gene expression levels can vary from one subtype to another, which results in multiple precision matrices for estimation. A naive approach is to model each subtype separately. However, in this separate approach, modeling of one subtype completely ignores the information on other subtypes. This can be suboptimal if there exists some common structure across different subtypes.

To improve the estimation in presence of some common structure, Guo et al. [2011]

proposed a joint estimation method in a penalized likelihood framework. This method employs a hierarchical penalty in the Gaussian likelihood framework to link the estimation of separate precision matrices. Their approach explores the common and unique structures via the hierarchical penalty.

In this chapter, we propose a new method to jointly estimate multiple precision matrices. Our approach uses a novel representation of each precision matrix as a sum of a common and unique matrices. Then we apply sparse constrained optimization on the common and unique components. The proposed method is applicable for a broader class of distributions including both the Gaussian and some non-Gaussian cases. The main strength of our method is that it utilizes all available information to jointly estimate the common and unique structures, which is not achievable in separate modelings. Therefore, the estimation can be improved if the precision matrices are similar to each other. Furthermore, our method is able to discover unique structures of each precision matrix, which enables us to identify differences among multiple precision matrices. The proposed estimator is shown to achieve a faster convergence rate for the common structure in certain cases.

The rest of this chapter is organized as follows. In Section 4.2, we introduce our proposed method after reviewing some separate approaches. We establish its theoretical properties in Section 4.3. Section 4.4 develops computational algorithms to obtain a solution for the proposed method. Simulated examples are presented in Section 4.5 to demonstrate performance of our estimator and analysis of a glioblastoma cancer data example is provided in Section 4.6. Section 4.7 provides some discussions. The proofs of theorems are provided in 4.8.

4.2 Methodology

In this section, we introduce a new method for estimating multiple precision matrices in a L_1 minimization framework. Consider a heterogeneous dataset with G different groups. For the gth group ($g = 1, \ldots, G$), let $\{\mathbf{x}_1^{(g)}, \ldots, \mathbf{x}_{n_g}^{(g)}\}$ be an independent and identically distributed random sample of size n_g, where $\mathbf{x}_k^{(g)} = (x_{ki}^{(g)}, \ldots, x_{kp}^{(g)})^{\mathrm{T}}$ is a p-dimensional random vector with the covariance matrix $\Sigma_0^{(g)}$ and precision matrix $\Omega_0^{(g)} := (\Sigma_0^{(g)})^{-1}$. For detailed illustration of our proposed method, we first define some notations similar to Cai, Liu and

Luo [2011]. For a matrix $X = (x_{ij}) \in \mathcal{R}^{p\times q}$, we define the elementwise L_1 norm $\|X\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{q}|x_{ij}|$, the elementwise l_∞ norm $|X|_\infty = \max_{1\le i\le p, 1\le j\le q}|x_{ij}|$ and the matrix L_1 norm $\|X\|_{L_1} = \max_{1\le j\le q}\sum_{i=1}^{p}|x_{ij}|$. For a vector $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathcal{R}^p$, $|x|_1$ and $|x|_\infty$ denote vector L_1 and l_∞ norm respectively. The notation $X \succ 0$ indicates that X is positive definite. Let I be a $p \times p$ identity matrix. For the gth group, $\hat{\Sigma}^{(\mathrm{g})}$ denotes the sample covariance matrix. Write $\Omega_0^{(\mathrm{g})} = (\omega_{ij,0}^{(\mathrm{g})}); g = 1, \ldots, G$.

Our aim is to estimate the precision matrices, $\Omega_0^{(1)}, \ldots, \Omega_0^{(\mathrm{G})}$. The most naive way to achieve this goal is to estimate each precision matrix separately by taking the inverses of the sample covariance matrices. However, in high dimensional cases, the sample covariance matrices are not only unstable for estimating the covariance matrices, but also not invertible. To estimate the precision matrix in high dimensions, various estimators have been introduced in the literature. For example, various L_1 penalized Gaussian likelihood estimators have been studied intensively in the literature [Yuan and Lin, 2007; Banerjee, Ghaoui and d'Aspremont, 2008; Friedman, Hastie and Tibshirani, 2008; Rothman et al., 2008]. In this framework, the precision matrices can be estimated by solving the following G optimization problems:

$$\min_{\Omega^{(\mathrm{g})}\succ 0} \mathrm{tr}(\hat{\Sigma}^{(\mathrm{g})}\Omega^{(\mathrm{g})}) - \log\{\det(\Omega^{(\mathrm{g})})\} + \lambda_g \sum_{i\neq j}|w_{ij}^{(\mathrm{g})}|, \ g = 1, \ldots, G, \tag{4.1}$$

where λ_g is a tuning parameter which controls the degree of the sparsity in the estimated precision matrices. Other sparse penalized Gaussian likelihood estimators have been proposed as well [Lam and Fan, 2009; Fan, Feng and Wu, 2009].

Recently, Cai, Liu and Luo [2011] proposed an interesting method of constrained L_1 minimization for inverse matrix estimation (CLIME), which can be directly implemented using linear programming. In particular, the CLIME estimator of $\Omega_0^{(\mathrm{g})}$ is the solution of the following optimization problem:

$$\min \|\Omega^{(\mathrm{g})}\|_1 \text{ subject to: } |\hat{\Sigma}^{(\mathrm{g})}\Omega^{(\mathrm{g})} - I|_\infty \le \lambda_g, \tag{4.2}$$

where $\hat{\Sigma}^{(\mathrm{g})}$ is the sample covariance matrix and λ_g is a tuning parameter. As the optimization problem in (4.2) does not require symmetry of the solution, the final CLIME estimator

is obtained by symmetrizing the solution of (4.2). The CLIME estimator does not need the Gaussian distributional assumption. Cai, Liu and Luo [2011] showed that the convergence rate of the CLIME estimator is faster than that of the L_1 penalized Gaussian likelihood estimator if the underlying true distribution has polynomial-type tails.

To estimate multiple precision matrices, $\Omega_0^{(1)}, \ldots, \Omega_0^{(G)}$, we can build G individual models using the optimization problem (4.1) or (4.2). However, these separate approaches can be suboptimal when the precision matrices share some common structure. For example, Guo et al. [2011] proposed a joint estimation of multiple precision matrices under the Gaussian distributional assumption to improve estimation. In particular, the estimator is the solution of

$$\min_{(\Omega^{(g)})_{g=1}^G} \sum_{g=1}^{G} \left[\mathrm{tr}(\hat{\Sigma}^{(g)}\Omega^{(g)}) - \log\{\det(\Omega^{(g)})\} \right] + \lambda_n \sum_{i \neq j} \left(\sum_{g=1}^{G} |\omega_{ij}^{(g)}| \right)^{1/2}, \tag{4.3}$$

where λ_n is a tuning parameter. In some simulation settings, they showed that the joint estimation can perform better than separate L_1 penalized normal likelihood estimation. However, this approach requires the Gaussian distributional assumption. In this chapter, we propose a new joint estimation of multiple precision matrices for both Gaussian and non-Gaussian cases.

In our joint estimation method, we first define the common structure M_0 and the unique structure $R_0^{(g)}$ as

$$M_0 := \frac{1}{G} \sum_{g=1}^{G} \Omega_0^{(g)}, R_0^{(g)} := \Omega_0^{(g)} - M_0; g = 1, \ldots, G.$$

It follows from the definition that $\sum_{g=1}^{G} R_0^{(g)} = 0$. If all precision matrices are very similar, then the unique structures defined above would be close to zero. In this case, it can be natural and advantageous to encourage sparsity among $\{R_0^{(1)}, \ldots, R_0^{(G)}\}$ in the estimation. To estimate the precision matrices consistently in high dimensions, it is also necessary to assume some special structure of M_0 as well. In our work, we also assume that M_0 is sparse. To estimate $\{M_0, R_0^{(1)}, \ldots, R_0^{(G)}\}$, we propose the following constrained L_1 minimization

criterion:

$$\min\{\|M\|_1 + \nu \sum_{g=1}^{G} \|R^{(g)}\|_1\}$$
$$\text{s.t } |\frac{1}{G}\sum_{g=1}^{G}\{\hat{\Sigma}^{(g)}(M + R^{(g)}) - I\}|_\infty \le \lambda_1, |\hat{\Sigma}^{(g)}(M + R^{(g)}) - I|_\infty \le \lambda_2, \sum_{g=1}^{G} R^{(g)} = 0, \tag{4.4}$$

where λ_1 and λ_2 are tuning parameters and ν is a prespecified weight. Note that if $\lambda_1 > \lambda_2$, then the second inequality constraints in (4.4) imply the first inequality constraint. Therefore, we only consider a pair of (λ_1, λ_2) satisfying $\lambda_1 \le \lambda_2$. The first inequality constraint in (4.4) reflects how close the final estimators are to the inverses of the sample covariance matrices in an average sense. On the other hand, the second inequality constraint controls an individual level of closeness between the estimators and the sample covariance matrices.

For illustration, consider an extreme case where all the precision matrices are the same. In this case, the unique structures can be negligible and the first inequality constraint in (4.4) reduces to $|(G^{-1}\sum_{g=1}^{G}\hat{\Sigma}^{(g)})M - I|_\infty \le \lambda_1$. Therefore, we can pool all the sample covariance matrices to estimate the common structure which is the precision matrix in this case. This would be advantageous than building each model separately. The value of ν in (4.4) reflects how complex the unique structures of the resulting estimators are. If the resulting estimators are expected to be very similar from each other, then a large value of ν is preferred. In Section 4.3, ν is set to be G^{-1} or $G^{-1/2}$ for our theoretical results.

Similar to Cai, Liu and Luo [2011], the solutions in (4.4) are not symmetric in general. Therefore, the final estimators are obtained after a symmetrization step. Denote the solution of (4.4) by $\{\hat{M}, \hat{R}^{(1)}, \ldots, \hat{R}^{(G)}\}$. Then we define $\hat{\Omega}_1^{(g)} := \hat{M} + \hat{R}^{(g)}; g = 1, \ldots, G$. The final estimator of $\{\Omega_0^{(1)}, \ldots, \Omega_0^{(G)}\}$ is obtained by symmetrizing $\{\hat{\Omega}_1^{(1)}, \ldots, \hat{\Omega}_1^{(G)}\}$ as follows. Let $\hat{\Omega}_1^{(g)} = (\hat{\omega}_{ij,1}^{(g)})$. Our joint estimator of multiple precision matrices (JEMP), $\{\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(G)}\}$, is defined as symmetric matrices, $\{\hat{\Omega}^{(g)} = (\hat{\omega}_{ij}^{(g)}); g = 1, \ldots, G\}$ with

$$\hat{\omega}_{ij}^{(g)} = \hat{\omega}_{ij,1}^{(g)} I\{\sum_{g=1}^{G}|\hat{\omega}_{ij,1}^{(g)}| \le \sum_{g=1}^{G}|\hat{\omega}_{ji,1}^{(g)}|\} + \hat{\omega}_{ji,1}^{(g)} I\{\sum_{g=1}^{G}|\hat{\omega}_{ij,1}^{(g)}| > \sum_{g=1}^{G}|\hat{\omega}_{ji,1}^{(g)}|\}; g = 1, \ldots, G. \tag{4.5}$$

4.3 Theoretical Properties

In this section, we investigate theoretical properties of our proposed joint estimator JEMP. In particular, we first construct the convergence rate of our estimator in the high dimensional setting. Then we show that the convergence rate can be improved for the common structure of the precision matrices in certain cases. Finally, the model selection consistency is shown with an additional thresholding step.

For theoretical properties, we follow the set-up of Cai, Liu and Luo [2011] and the results therein are also used for our technical derivations. In this section, for simplicity, we assume that $n = n_1 = \cdots = n_G$. We consider the following class of matrices,

$$\mathcal{U} := \{\Omega : \Omega \succ 0, \|\Omega\|_{L_1} \leq C_M\},$$

and assume that $\Omega_0^{(g)} \in \mathcal{U}$ for all $g = 1, \ldots, G$. This assumption requires that the true precision matrices are sparse in terms of the L_1 norm while allowing them to have many small entries. Write $E(\mathbf{x}^{(g)}) = (\mu_1^{(g)}, \ldots, \mu_p^{(g)})^{\mathrm{T}}$. We also make the following moment condition on $\mathbf{x}^{(g)}$ for our theoretical results.

Condition 1. *There exists some $0 < \eta < 1/4$ such that $E[\exp\{t(x_i^{(g)} - \mu_i^{(g)})^2\}] \leq K < \infty$ for all $|t| \leq \eta$ and all i, g and $G \log p/n \leq \eta$, where K is a bounded constant.*

Condition 1 indicates that the components of $\mathbf{x}^{(g)}$ are uniformly sub-Gaussian. This condition is satisfied if $\mathbf{x}^{(g)}$ follows a multivariate Gaussian distribution or has uniformly bounded components.

Theorem 7. *Assume Condition 1 holds. Let $\lambda_1 = \lambda_2 = 3C_M C_0 (\log p/n)^{1/2}$, where $C_0 = 2\eta^{-2}(2 + \tau + \eta^{-1}e^2K^2)^2$ and $\tau > 0$. Set $\nu = G^{-1}$. Then*

$$\max_{ij}\left(\frac{1}{G}\sum_{g=1}^{G}|\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}|\right) \leq 6C_M^2 C_0\left(\frac{\log p}{n}\right)^{1/2}, \tag{4.6}$$

with probability greater than $1 - 4Gp^{-\tau}$.

In an average sense, the convergence rate can be viewed the same as that of the CLIME estimator which is of order $(\log p/n)^{1/2}$. In this theorem, the first inequality constraint in (4.4) does not play any role in the estimation procedure as we set $\lambda_1 = \lambda_2$. In the next

theorem, with properly chosen λ_1, we construct a faster convergence rate for the common part under certain conditions.

Theorem 8. *Assume Condition 1 holds. Suppose that there exists $C_R > 0$ such that $\|R_0^{(g)}\|_{L_1} \le C_R$ for all $g = 1, \ldots, G$ and $(\sum_{g=1}^{G} \|R_0^{(g)}\|_{L_1}) \le C_R G^{1/2}$. Set $\nu = G^{-1/2}$ and let $\lambda_1 = (C_M + C_R)C_0\{\log p/(nG)\}^{1/2}$ and $\lambda_2 = C_M C_0 (\log p/n)^{1/2}$. Then*

$$|\hat{M} - M_0|_\infty \le C_0(2C_M^2 + 4C_M C_R + C_R^2)\left(\frac{\log p}{nG}\right)^{1/2}, \tag{4.7}$$

with probability greater than $1 - 2(1 + 3G)p^{-\tau}$.

Theorem 8 states that our proposed method can estimate the common part more efficiently with the corresponding convergence rate of order $\{\log p/(nG)\}^{1/2}$, which is faster than the order $(\log p/n)^{1/2}$.

Besides its estimation consistency, we also prove the model selection consistency of our estimator which means that it reveals the exact set of nonzero components in the true precision matrices with high probability. For this result, a thresholding step is introduced. In particular, a threshold estimator $\tilde{\Omega}^{(g)} = (\tilde{\omega}_{ij}^{(g)})$ based on $\{\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(G)}\}$ is defined as,

$$\tilde{\omega}_{ij}^{(g)} = \hat{\omega}_{ij}^{(g)} I\{|\hat{\omega}_{ij}^{(g)}| \ge \delta_n\}, \tag{4.8}$$

where $\delta_n \ge 2C_M G\lambda_2$ and λ_2 is given in Theorem 7. To state the model selection consistency precisely, we define

$$\mathcal{S}_0 := \{(i,j,g) : \omega_{ij,0}^{(g)} \ne 0\}, \hat{\mathcal{S}} := \{(i,j,g) : \tilde{\omega}_{ij}^{(g)} \ne 0\} \text{ and } \theta_{\min} := \min_{(i,j)\in\mathcal{S}_0} \sum_{g=1}^{G} |\omega_{ij,0}^{(g)}|.$$

Then the next theorem states the model selection consistency of our estimator.

Theorem 9. *Assume Condition 1 holds. If $\theta_{\min} > 2\delta_n$ and $Gp^{-\tau} \to 0$, then*

$$pr(\mathcal{S}_0 = \hat{\mathcal{S}}) \to 1.$$

4.4 Numerical Algorithm

In this section, we describe how to obtain the numerical solutions of the optimization problem (4.4). Similar to the Lemma 1 in Cai, Liu and Luo [2011], we can show that the optimization problem (4.4) can be decomposed into p individual minimization problems. In particular, let e_i be the ith column of I. For $1 \le i \le p$, let $\{\hat{m}_i, \hat{r}_i^{(1)}, \ldots, \hat{r}_i^{(G)}\}$ be the solution of the following optimization problem:

$$\min\{|m|_1 + \nu \sum_{g=1}^{G} |r^{(g)}|_1\}$$
$$\text{s.t. } |\frac{1}{G}\sum_{g=1}^{G}\{\hat{\Sigma}^{(g)}(m + r^{(g)}) - e_i\}|_\infty \le \lambda_1, |\hat{\Sigma}^{(g)}(m + r^{(g)}) - e_i|_\infty \le \lambda_2, \sum_{g=1}^{G} r^{(g)} = 0, \tag{4.9}$$

where $m, r^{(1)}, \ldots, r^{(G)}$ are vectors in $\mathcal{R}^p$. We can show that solving the optimization problem (4.4) is equivalent to solving the p optimization problems in (4.9). The optimization problem in (4.9) can be further relaxed to be a linear programming problem and the simplex method approach is used to solve this problem [Boyd and Vandenberghe, 2004].

To apply our method, we need to choose the tuning parameters, λ_1 and λ_2. In practice, we construct several models with many pairs of λ_1 and λ_2 satisfying $\lambda_1 \le \lambda_2$ and evaluate them to determine the optimal pair. To evaluate each estimator, we measure the likelihood loss (LL) used in Cai, Liu and Luo [2011] and its definition is

$$\text{LL} = \sum_{g=1}^{G} \text{tr}(\hat{\Sigma}_v^{(g)}\hat{\Omega}^{(g)}) - \log\{\det(\hat{\Omega}^{(g)})\},$$

where $\hat{\Sigma}_v^{(g)}$ is the sample covariance matrix of the gth group computed from an independent validation set. Among several pairs of tuning values, we select the pair which minimizes LL. If a validation set is not available, a K-fold cross-validation can be combined to this criterion. In particular, we first randomly split the dataset into K parts of equal sizes. Denote the data in the kth part by $\{X_{(k)}^{(1)}, \ldots, X_{(k)}^{(G)}\}$ which is used as a validation set for the kth estimator. For each k, with a given value of (λ_1, λ_2), we obtain an estimator using all observations which do not belong to $\{X_{(k)}^{(1)}, \ldots, X_{(k)}^{(G)}\}$ and denote them as $\{\hat{\Omega}_{(k)}^{(G)}, \ldots, \hat{\Omega}_{(k)}^{(G)}\}$.

Then the likelihood loss (LL) is defined as

$$\mathrm{LL} = \sum_{k=1}^{K} \sum_{g=1}^{G} \mathrm{tr}(\hat{\Sigma}_{(k)}^{(g)} \hat{\Omega}_{(k)}^{(g)}) - \log\{\det(\hat{\Omega}_{(k)}^{(g)})\},$$

where $\hat{\Sigma}_{(k)}^{(g)}$ is the sample covariance matrix of the gth group using $X_{(k)}^{(g)}$. Once the optimal pair is selected which minimizes LL, the final model is constructed using all data points with the selected pair.

4.5 Simulated Examples

In this section, we carry out simulation studies to assess the numerical performance of our proposed method. In particular, we compare the numerical performance of four methods: two separate methods and two joint methods. In separate approaches, each precision matrix is estimated separately via the CLIME estimator or the GLASSO estimator. On the other hand, in joint approaches, all precision matrices are estimated together using our JEMP estimator or the joint estimator by Guo et al. [2011], which we refer to as JOINT estimator hereafter. In our proposed method, ν is set to be $G^{-1/2}$. We also tried different values of ν such as G^{-1}, and the results are similar thus omitted. We consider three models as described below: the first two from Guo et al. [2011] and the last from Rothman et al. [2008]; Cai, Liu and Luo [2011]. In all models, we set $p = 100$, $G = 3$ and $\Omega_0^{(g)} = \Omega_c + U^{(g)}$, where Ω_c is common in all groups and $U^{(g)}$ represents unique structure to the gth group. The common part, Ω_c, is generated as follows:

Model 1. Ω_c is a tridiagonal precision matrix. In particular, $\Sigma_c := \Omega_c^{-1} = (\sigma_{ij})$ is first constructed, where $\sigma_{ij} = \exp(-|d_i - d_j|/2)$, $d_1 < \ldots < d_p$, and $d_i - d_{i-1} \sim \mathrm{Unif}(0.5, 1), i = 2, \ldots, p$. Then let $\Omega_c = \Sigma_c^{-1}$.

Model 2. Ω_c is a 5 nearest-neighbor network. In particular, p points are randomly picked on a unit square and all pairwise distances among the points are calculated. Then we find 5 nearest neighbors for each point and a pair of symmetric entries in Ω_c corresponding to a pair of neighbors has a value randomly chosen from the interval $[-1, -0.5] \cup [0.5, 1]$.

Model 3. $\Omega_c = \Gamma + \delta I$, where each off-diagonal entry in Γ is generated independently from $0.5y$, with y following the Bernoulli distribution with success probability 0.1. Here, δ is selected so that the condition number of Γ is equal to p.

For each $U^{(g)}$, we randomly pick a pair of symmetric off-diagonal entries and replace them with values randomly chosen from the interval $[-1,-0.5] \cup [0.5,1]$. We repeat this procedure until $\sum_{i<j} I(|u_{ij}^{(g)}| > 0) / \sum_{i<j} I(|\omega_{ij,c}| > 0) = \rho$, where $\Omega_c = (\omega_{ij,c})$ and $U^{(g)} = u_{ij}^{(g)}$. Therefore, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries. We consider four values of $\rho = 0, 0.25, 1$ and 4. Finally, each matrix $\Omega_0^{(g)}$ is standardized to have unit diagonals.

For each group in each model, we generate a training sample of size $n = 100$ from a multivariate normal distribution $N(0, \Sigma_0^{(g)})$. In order to select optimal tuning parameters, an independent validation set of size $n = 100$ is also generated from the same distribution. For each estimator, optimal tuning parameters are selected as described in Section 4.4. We replicate simulations 50 times for each model. To compare performance of four different methods, we use several criteria as follows. For estimation quality, we use the average entropy loss and the average Frobenius loss defined as,

$$\text{EL} = G^{-1} \sum_{g=1}^{G} \left\{ \text{tr}(\Sigma_0^{(g)} \hat{\Omega}^{(g)}) - \log\det(\Sigma_0^{(g)} \hat{\Omega}^{(g)}) - p \right\},$$
$$\text{FL} = G^{-1} \sum_{g=1}^{G} \| \Omega_0^{(g)} - \hat{\Omega}^{(g)} \|_F^2,$$

where $\| \cdot \|_F$ is the Frobenius norm of a matrix. To measure selection quality, we use the average false positive rate and the average false negative rate defined as,

$$\text{FP} = \frac{1}{G} \sum_{g=1}^{G} \frac{\sum_{i<j} I(\omega_{ij,0}^{(g)} = 0, \hat{\omega}_{ij}^{(g)} \neq 0)}{\sum_{i<j} I(\omega_{ij,0}^{(g)} = 0)},$$
$$\text{FN} = \frac{1}{G} \sum_{g=1}^{G} \frac{\sum_{i<j} I(\omega_{ij,0}^{(g)} \neq 0, \hat{\omega}_{ij}^{(g)} = 0)}{\sum_{i<j} I(\omega_{ij,0}^{(g)} \neq 0)}.$$

Table 4.1 reports the results for all models. In terms of the average entropy loss and Frobenius loss, two joint estimation methods, JEMP and JOINT, outperform two separate estimation methods. Our proposed method, JEMP, has the best performance overall. JEMP shows slightly worse false positive rates. On the other hand, from all examples, it can be seen that there are some improvements in the false negative rates.

Figure 4.1 shows the estimated receiver operating characteristic (ROC) curves averaged over 50 replications. In Model 1, the ROC curves estimated by JEMP and JOINT seem

Table 4.1: Average entropy loss (EL), Frobenius loss (FL), false positive rate (FP), and false negative rate (FN) for three models over 50 replications (The numbers in parentheses are standard errors)

Model	ρ	Method	EL	FL	FP	FN
Model 1	0	CLIME	7.58 (0.035)	12.69 (0.066)	0.05 (0.000)	0.05 (0.002)
		GLASSO	6.65 (0.027)	10.79 (0.045)	0.09 (0.001)	0.03 (0.001)
		JOINT	3.03 (0.025)	4.87 (0.044)	0.00 (0.000)	0.01 (0.001)
		JEMP	2.41 (0.020)	5.04 (0.060)	0.09 (0.005)	0.00 (0.000)
	0.25	CLIME	7.60 (0.032)	12.97 (0.053)	0.05 (0.000)	0.23 (0.002)
		GLASSO	6.81 (0.034)	10.99 (0.048)	0.09 (0.001)	0.18 (0.002)
		JOINT	3.90 (0.031)	6.16 (0.043)	0.00 (0.000)	0.21 (0.003)
		JEMP	3.13 (0.021)	6.03 (0.065)	0.10 (0.006)	0.16 (0.002)
	1	CLIME	7.09 (0.028)	12.72 (0.064)	0.05 (0.001)	0.56 (0.003)
		GLASSO	6.52 (0.024)	10.88 (0.042)	0.07 (0.001)	0.54 (0.004)
		JOINT	5.19 (0.032)	8.58 (0.057)	0.00 (0.000)	0.63 (0.004)
		JEMP	3.87 (0.022)	7.39 (0.065)	0.10 (0.006)	0.43 (0.005)
	4	CLIME	4.75 (0.018)	9.18 (0.035)	0.04 (0.001)	0.90 (0.003)
		GLASSO	3.97 (0.012)	7.29 (0.024)	0.01 (0.001)	0.96 (0.003)
		JOINT	3.97 (0.012)	7.33 (0.024)	0.00 (0.000)	0.99 (0.001)
		JEMP	3.11 (0.013)	6.28 (0.028)	0.07 (0.001)	0.82 (0.001)
Model 2	0	CLIME	7.06 (0..027)	12.93 (0.056)	0.05 (0.001)	0.58 (0.003)
		GLASSO	6.45 (0.023)	11.06 (0.034)	0.06 (0.001)	0.55 (0.004)
		JOINT	5.57 (0.034)	9.86 (0.053)	0.00 (0.000)	0.66 (0.007)
		JEMP	3.27 (0.024)	6.29 (0.079)	0.15 (0.006)	0.14 (0.007)
	0.25	CLIME	6.33 (0.025)	11.66 (0.062)	0.05 (0.001)	0.70 (0.003)
		GLASSO	5.78 (0.022)	9.93 (0.035)	0.05 (0.001)	0.69 (0.004)
		JOINT	6.00 (0.030)	10.33 (0.045)	0.01 (0.000)	0.84 (0.004)
		JEMP	5.47 (0.019)	10.80 (0.033)	0.12 (0.006)	0.48 (0.007)
	1	CLIME	5.46 (0.020)	10.44 (0.045)	0.05 (0.001)	0.85 (0.003)
		GLASSO	4.85 (0.015)	8.66 (0.027)	0.03 (0.001)	0.89 (0.004)
		JOINT	5.09 (0.017)	9.19 (0.032)	0.00 (0.000)	0.99 (0.001)
		JEMP	4.73 (0.014)	9.45 (0.031)	0.09 (0.002)	0.83 (0.004)
	4	CLIME	3.47 (0.017)	6.62 (0.031)	0.04 (0.001)	0.95 (0.001)
		GLASSO	2.45 (0.013)	4.74 (0.027)	0.00 (0.000)	1.00 (0.000)
		JOINT	2.45 (0.012)	4.73 (0.026)	0.00 (0.000)	1.00 (0.000)
		JEMP	2.17 (0.011)	4.37 (0.023)	0.07 (0.001)	0.93 (0.001)
Model 3	0	CLIME	4.21 (0.019)	8.27 (0.034)	0.04 (0.001)	0.92 (0.002)
		GLASSO	3.29 (0.013)	6.30 (0.026)	0.00 (0.001)	0.99 (0.001)
		JOINT	3.28 (0.014)	6.30 (0.027)	0.00 (0.000)	1.00 (0.000)
		JEMP	2.68 (0.010)	5.57 (0.023)	0.07 (0.001)	0.77 (0.002)
	0.25	CLIME	4.55 (0.017)	8.83 (0.034)	0.04 (0.001)	0.91 (0.002)
		GLASSO	3.78 (0.014)	7.04 (0.026)	0.01 (0.001)	0.97 (0.002)
		JOINT	3.85 (0.014)	7.19 (0.026)	0.00 (0.000)	1.00 (0.000)
		JEMP	3.47 (0.010)	7.05 (0.023)	0.07 (0.001)	0.88 (0.002)
	1	CLIME	3.68 (0.014)	7.06 (0.026)	0.04 (0.001)	0.95 (0.001)
		GLASSO	2.69 (0.013)	5.15 (0.027)	0.00 (0.000)	1.00 (0.001)
		JOINT	2.69 (0.013)	5.15 (0.027)	0.00 (0.000)	1.00 (0.000)
		JEMP	2.42 (0.009)	4.87 (0.020)	0.07 (0.001)	0.92 (0.001)
	4	CLIME	2.67 (0.018)	4.92 (0.036)	0.04 (0.001)	0.96 (0.001)
		GLASSO	1.60 (0.015)	3.27 (0.035)	0.00 (0.000)	1.00 (0.000)
		JOINT	1.60 (0.015)	3.26 (0.036)	0.00 (0.000)	1.00 (0.000)
		JEMP	1.36 (0.010)	2.68 (0.018)	0.07 (0.001)	0.93 (0.001)

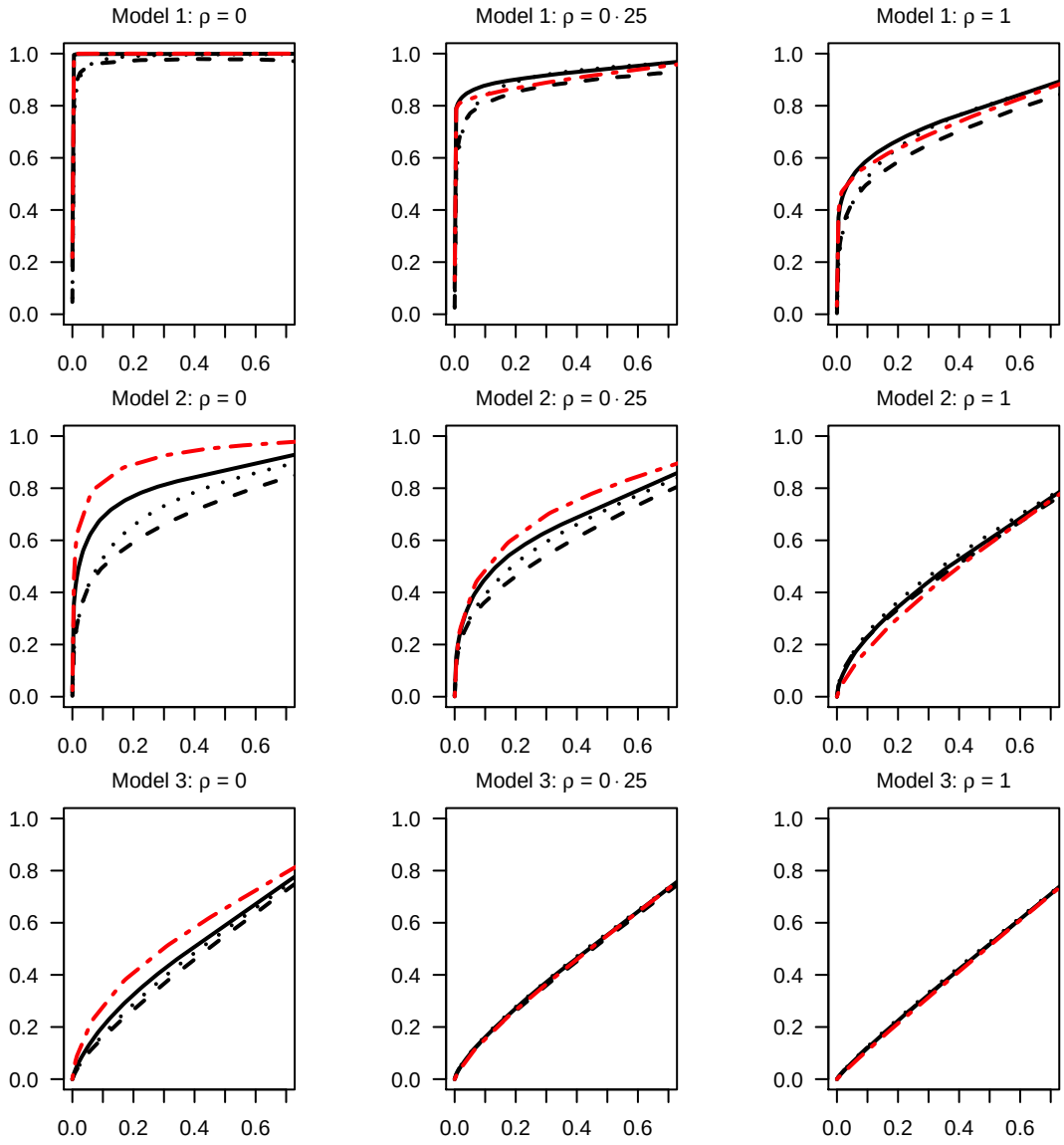

Figure 4.1: Receiver operating characteristic curves averaged over 50 replications. In each panel, the horizontal and vertical axes are false positive rate and sensitivity respectively. Here, ρ is the ratio of the number of unique nonzero entries to the number of common nonzero entries. The red dotted-dash, solid, dotted, and dashed lines correspond to JEMP, JOINT, GLASSO, and CLIME respectively.

to be close while they dominate the ROC curves estimated by CLIME and GLASSO. In Model 2, when ρ is 0 or 0.25, the ROC curve from JEMP dominates the other curves. As ρ increases, all ROC curves move closer together. This is because the precision matrices become much different from each other and the joint estimation methods eventually do not have any advantage. In Model 3, JEMP estimator outperforms the other estimators when ρ is 0. Overall, our proposed JEMP estimator delivers very competitive performance in terms of both estimation accuracy and selection.

4.6 Application to the Glioblastoma Cancer Data

In this section, we apply our methodology to the GBM cancer dataset described in Section 3.6. The dataset consists of 17814 gene expression levels of 482 GBM patients. The patients were classified into four subtypes, namely, classical, mesenchymal, neural, and proneural with sample sizes of 127, 145, 85, and 125 respectively [Verhaak et al., 2010]. These subtypes are shown to be different biologically, while at the same time, share similarities as well since they all belong to GBM cancer. In this application, we consider the signature genes reported by Verhaak et al. [2010]. They established 210 signature genes for each subtype, which results 840 signature genes in total. These signature genes are highly distinctive for four subtypes and each class of genes tends to be highly expressed only in their corresponding subtype.

To produce an interpretable size of models, top 15 genes with large median absolute deviation were selected from each set of 210 signature genes, which results 60 genes in total. Our aim is to estimate four precision matrices of the 60 genes for four corresponding subtypes. We consider four methods described in Section 4.5. For performance assessment, the dataset of each subtype is divided into a training set of size 75 and a test set of the remainder. The tuning parameters are selected using 5-fold cross-validation as discussed in Section 4.4. We perform the random splitting 100 times. Each estimator is evaluated using the likelihood loss on the test set defined in Section 4.4.

Table 4.2 shows the average likelihood loss based on 100 replications. The joint estimation methods outperform the separate approaches. Among all estimators, our estimator shows the lowest likelihood loss. This result indicates that there may exist some noticeable

Table 4.2: Comparison of the average likelihood loss based on 100 replications (The numbers in parentheses are standard errors)

	CLIME	GLASSO	JOINT	JEMP
Likelihood loss	91.71(0.39)	87.03(0.66)	74.80(0.94)	67.01(0.36)

common structure shared by all precision matrices.

To depict the common and unique structures among the estimated precision matrices, graphical networks are constructed using our joint estimator. In each subtype, two genes have an edge if the corresponding element in the estimated precision matrix is nonzero for all 100 replications. The resulting gene networks are shown in Figure 4.2. The thin dark grey lines are the edges appearing in all subtypes and the thick black lines are the unique edges to certain subtypes. It is noticeable that most of edges are dark grey lines, which means that they appear in all subtypes. This indicates that the networks of 60 genes from GBM patients are very similar across all subtypes although they have some unique structures for each subtype as well. The red genes are signature genes for the classical subtype. Likewise, green, blue and purple genes are the mesenchymal, proneural and neural signature genes respectively. Each class of signature genes tends to have more links with the genes in the same class. This is expected because each class of signature genes is more likely to be highly co-expressed. Some genes such as GJA1, ELOVL2, and FHL2, have many links with other class of genes. This may indicate that these genes may have information for multiple subtypes and it will be interesting to investigate these genes further.

4.7 Discussion

In this chapter, we proposed a new method for jointly estimating multiple precision matrices with some common structure. The proposed method is derived in a constrained L_1 minimization framework. Our theoretical investigation shows that the estimation can be improved for the common structure in certain cases. Simulated examples and an application to the GBM cancer data set demonstrate that our proposed methods perform competitively.

Our current method defines the common structure of multiple precision matrices as the average of these matrices. One future research direction is to extend the proposed method

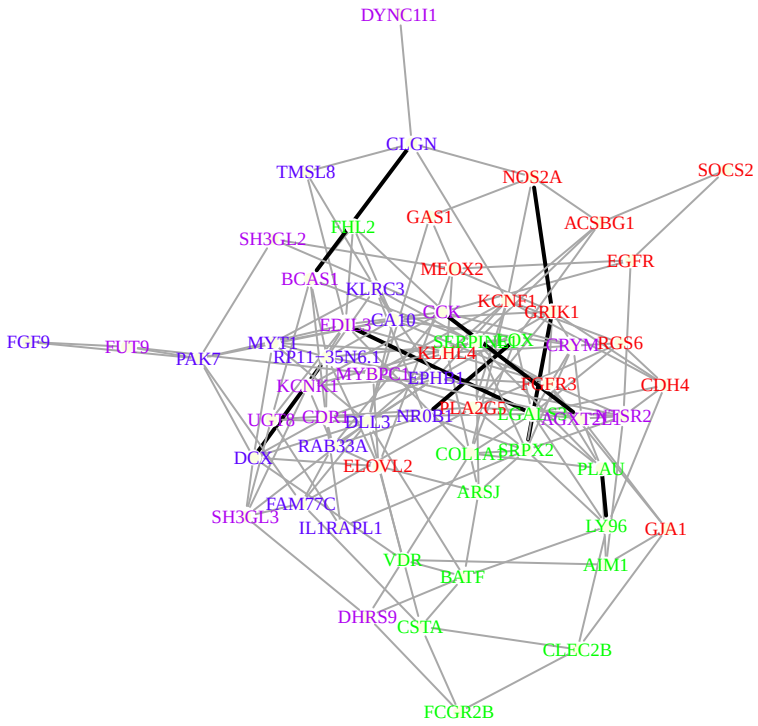

Figure 4.2: Graphical presentation of conditional dependence structures among genes using our estimator of precision matrices. The thin dark grey lines are the edges appearing in all subtypes and the thick black lines are the unique edges to certain subtypes. The red, green, blue and purple genes are classical, mesenchymal, proneural and neural genes respectively [Verhaak et al., 2010].

with other definitions of the common structure. For instance, the common structure can be defined as the intersection of the index sets of non-zero components of the precision matrices. It will be interesting to compare the performance with various definitions of the common structure.

Our methodology mainly focuses on the sparse estimation of precision matrices. The proposed method can be very useful statistical tools for exploring common and unique structures of multiple precision matrices. On the other hand, it is not clear how to make valid inferences on our estimator. Although some parameters are estimated numerically as zeros by our sparse technique, one may need to perform hypothesis tests to determine if they are zeros in the statistical sense. Therefore, a natural future direction is to develop valid inference tools for our estimator, such as performing hypothesis tests and constructing confidence intervals.

4.8 Proofs

4.8.1 Proof of Theorem 7

Write $\Sigma_0^{(g)} = (\sigma_{ij,0}^{(g)})$ and $\hat{\Sigma}^{(g)} = (\hat{\sigma}_{ij}^{(g)})$. Let $m_{j,0}$ and $r_{j,0}^{(g)}$ be the jth columns of M_0 and $R_0^{(g)}$ respectively. Define the jth columns of $\hat{M}$ and $\hat{R}^{(g)}$ as $\hat{m}_j$ and $\hat{r}_j^{(g)}$ respectively. We first state some results established by Cai, Liu and Luo [2011] in the proof of their Theorem 1.

lemma 4. *Suppose Condition 1 holds. For any fixed $g = 1, \ldots, G$, with probability greater than $1 - 4p^{-\tau}$,*

$$\max_{ij} |\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)}| \leq C_0 \left(\frac{\log p}{n} \right)^{1/2},$$

where C_0 is given in Theorem 7.

It follows from Lemma 4 that

$$\max_{ij} |\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)}| \leq \lambda_2/(3C_M) \quad \text{for all } g = 1, \ldots, G, \tag{4.10}$$

with probability greater than $1-4Gp^{-\tau}$. All following arguments assume (4.10) holds. First, we have that

$$\begin{aligned}
|(\hat{\Omega}_1^{(g)} - \Omega_0^{(g)})e_j|_\infty &= |\Omega_0^{(g)}(\Sigma_0^{(g)}\hat{\Omega}_1^{(g)} - I)e_j|_\infty \leq \|\Omega_0^{(g)}\|_{L_1} |(\Sigma_0^{(g)}\hat{\Omega}_1^{(g)} - I)e_j|_\infty \\
&\leq C_M \left\{ |(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})\hat{\Omega}_1^{(g)} e_j|_\infty + |(\hat{\Sigma}^{(g)}\hat{\Omega}_1^{(g)} - I)e_j|_\infty \right\} \\
&\leq C_M |\hat{\Omega}_1^{(g)} e_j|_1 |\Sigma_0^{(g)} - \hat{\Sigma}^{(g)}|_\infty + C_M \lambda_2 \\
&\leq |\hat{\Omega}_1^{(g)} e_j|_1 \lambda_2/3 + C_M \lambda_2,
\end{aligned}$$

for all $g = 1, \ldots, G$. Second, note that $\{M_0, R_0^{(1)}, \ldots, R_0^{(G)}\}$ is a feasible solution of (4.4) as $|I - \hat{\Sigma}^{(g)}(M_0 + R_0^{(g)})|_\infty = |(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})\Omega_0^{(g)}|_\infty \leq \|\Omega_0^{(g)}\|_{L_1} |\Sigma_0^{(g)} - \hat{\Sigma}^{(g)}|_\infty \leq C_M \lambda_2/(3C_M) < \lambda_2$ and $\lambda_1 = \lambda_2$. Therefore, we have that

$$\begin{aligned}
\sum_{g=1}^{G} |(\hat{\Omega}_1^{(g)} - \Omega_0^{(g)})e_j|_\infty &\leq \sum_{g=1}^{G} |\hat{\Omega}_1^{(g)} e_j|_1 \lambda_2/3 + GC_M\lambda_2 \leq G \left\{ |\hat{m}_j|_1 + G^{-1} \sum_{g=1}^{G} |\hat{r}_j^{(g)}|_1 \right\} \lambda_2/3 + GC_M\lambda_2 \\
&\leq G \left\{ |m_{j,0}|_1 + G^{-1} \sum_{g=1}^{G} |r_{j,0}^{(g)}|_1 \right\} \lambda_2/3 + GC_M\lambda_2 \\
&\leq G3C_M\lambda_2/3 + GC_M\lambda_2 = 2GC_M\lambda_2 = 6GC_M^2 C_0 (\log p/n)^{1/2}.
\end{aligned}$$

By the inequality

$$\max_{ij}\left(\frac{1}{G}\sum_{g=1}^{G}|\hat{\omega}_{ij}^{(g)}-\omega_{ij,0}^{(g)}|\right)\leq\max_{j}\frac{1}{G}\sum_{g=1}^{G}|(\hat{\Omega}_{1}^{(g)}-\Omega_{0}^{(g)})e_{j}|_{\infty}\leq 6C_{M}^{2}C_{0}\Big(\frac{\log p}{n}\Big)^{1/2},$$

the proof is completed.

4.8.2 Proof of Theorem 8

lemma 5. *With probability greater than* $1-2(1+G)p^{-\tau}$*, the following holds:*

$$\max_{ij}|\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(g)}-\sigma_{ij,0}^{(g)})|\leq C_{0}\left(\frac{G\log p}{n}\right)^{1/2}.$$

Proof. We adopt a similar technique used in Cai, Liu and Luo [2011] for the proof of their Theorem 1. Without loss of generality, we assume that $\mu_i^{(g)}=0$ for all i and g. Let $y_{kij}^{(g)}:=x_{ki}^{(g)}x_{kj}^{(g)}-E(x_{ki}^{(g)}x_{kj}^{(g)})$. Define $\bar{x}_i^{(g)}:=\sum_{k=1}^{n}x_{ki}^{(g)}/n; i=1,\ldots,p, g=1,\ldots,G$. Then $\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(g)}-\sigma_{ij,0}^{(g)})=\sum_{g=1}^{G}\left(\sum_{k=1}^{n}y_{kij}^{(g)}/n-\bar{x}_i^{(g)}\bar{x}_j^{(g)}\right)$. Let $t:=\eta(\log p)^{1/2}(nG)^{-1/2}$ and $C_1:=2+\tau+\eta^{-1}K^2$. Using the Markov's inequality and the inequality $|\exp(s)-1-s|\leq s^2\exp\{\max(s,0)\}$ for any $s\in\mathcal{R}$, we can show that

$$\begin{aligned}
\mathrm{pr}\left\{\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n}y_{kij}^{(g)}\geq\eta^{-1}C_{1}\left(\frac{G\log p}{n}\right)^{1/2}\right\}&=\mathrm{pr}\left\{\sum_{g=1}^{G}\sum_{k=1}^{n}y_{kij}^{(g)}\geq\eta^{-1}C_{1}\left(nG\log p\right)^{1/2}\right\}\\
&\leq\exp\left\{-t\eta^{-1}C_{1}(nG\log p)^{1/2}\right\}E\left\{\exp\left(t\sum_{g=1}^{G}\sum_{k=1}^{n}y_{kij}^{(g)}\right)\right\}\\
&=\exp\left\{-C_{1}\log p\right\}\prod_{g=1}^{G}\prod_{k=1}^{n}E\left\{\exp(ty_{kij}^{(g)})\right\}\\
&=\exp\left[-C_{1}\log p+\sum_{g=1}^{G}n\log\left\{E\left(e^{ty_{kij}^{(g)}}\right)\right\}\right]\\
&\leq\exp\left[-C_{1}\log p+\sum_{g=1}^{G}n\left\{E\left(e^{ty_{kij}^{(g)}}\right)-1\right\}\right]\\
&=\exp\left[-C_{1}\log p+\sum_{g=1}^{G}n\left\{E\left(e^{ty_{kij}^{(g)}}-ty_{kij}^{(g)}-1\right)\right\}\right]\\
&\leq\exp\left\{-C_{1}\log p+\sum_{g=1}^{G}nt^{2}E\left({y_{kij}^{(g)}}^{2}e^{|ty_{kij}^{(g)}|}\right)\right\}\\
&\leq\exp\left\{-C_{1}\log p+\sum_{g=1}^{G}(\eta G)^{-1}K^{2}\log p\right\}. \qquad (4.11)
\end{aligned}$$

The last inequality (4.11) holds since $nt^2E\left({y_{kij}^{(g)}}^2e^{|ty_{kij}^{(g)}|}\right)=(\eta G)^{-1}(\log p)E\left\{\left(\eta^{3/2}|y_{kij}^{(g)}|\right)^2e^{t|y_{kij}^{(g)}|}\right\}$

and

$$E\left\{\left(\eta^{3/2}|y_{kij}^{(\mathrm{g})}|\right)^2 e^{t|y_{kij}^{(\mathrm{g})}|}\right\} \leq E\left\{e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}e^{t|y_{kij}^{(\mathrm{g})}|}\right\} \leq E\left\{e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}e^{\eta^{3/2}|y_{kij}^{(\mathrm{g})}|}\right\} \leq E\left\{e^{\eta|y_{kij}^{(\mathrm{g})}|}\right\}$$
$$\leq E\left\{e^{\eta|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}|+\eta E\left(|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}|\right)}\right\} \leq \left\{E\left(e^{\eta|x_{ki}^{(\mathrm{g})}x_{kj}^{(\mathrm{g})}|}\right)\right\}^2$$
$$\leq \left\{E\left(e^{\eta {x_{ki}^{(\mathrm{g})}}^2/2+\eta {x_{kj}^{(\mathrm{g})}}^2/2}\right)\right\}^2 \leq E\left(e^{\eta {x_{ki}^{(\mathrm{g})}}^2}\right)E\left(e^{\eta {x_{kj}^{(\mathrm{g})}}^2}\right) \leq K^2.$$

From (4.11), it follows that

$$\mathrm{pr}\left\{\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})} \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} \leq \exp\left\{-C_1\log p + \eta^{-1}K^2\log p\right\} \leq p^{-(\tau+2)}.$$

Therefore, we have

$$\mathrm{pr}\left\{\max_{ij}\left|\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(\mathrm{g})}\right| \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} \leq 2p^{-\tau}. \tag{4.12}$$

Next, let $C_2 = 2 + \tau + \eta^{-1}(eK)^2$. Cai, Liu and Luo [2011] showed in the proof of their Theorem 1 that

$$\mathrm{pr}\left(\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2\log p/n\right) \leq 2p^{-\tau-1}.$$

Using this result, we have that

$$\mathrm{pr}\left(\max_{ij}|\sum_{g=1}^{G}\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2 G\log p/n\right) \leq \mathrm{pr}\left(\sum_{g=1}^{G}\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2 G\log p/n\right)$$
$$\leq \sum_{g=1}^{G}\mathrm{pr}\left(\max_{ij}|\bar{x}_i^{(\mathrm{g})}\bar{x}_j^{(\mathrm{g})}| \geq \eta^{-2}C_2^2\log p/n\right)$$
$$\leq \sum_{g=1}^{G} 2p^{-\tau-1} \leq 2Gp^{-\tau} \tag{4.13}$$

By (4.12), (4.13) and the inequality $C_0 > \eta^{-1}C_1 + \eta^{-2}C_2^2(G\log p/n)^{1/2}$, we see that

$$\begin{aligned}
&\mathrm{pr}\left(\max_{ij}|\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)})| \geq C_0\left(\frac{G\log p}{n}\right)^{1/2}\right) \\
&\leq \mathrm{pr}\left\{\max_{ij}\left|\frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n} y_{kij}^{(g)}\right| \geq \eta^{-1}C_1\left(\frac{G\log p}{n}\right)^{1/2}\right\} + \mathrm{pr}\left(\max_{ij}|\sum_{g=1}^{G}\bar{x}_i^{(g)}\bar{x}_j^{(g)}| \geq \eta^{-2}C_2^2 G\log p/n\right) \\
&\leq 2(1+G)p^{-\tau}.
\end{aligned}$$

The proof is completed. □

By Lemma 4 and 5, we see that

$$\max_{ij}|\sum_{g=1}^{G}(\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)})| \leq C_0\left(\frac{G\log p}{n}\right)^{1/2} \text{and} \max_{ij}|\hat{\sigma}_{ij}^{(g)} - \sigma_{ij,0}^{(g)}| \leq C_0\left(\frac{\log p}{n}\right)^{1/2}, \tag{4.14}$$

for all $g = 1,\ldots,G$ with probability greater than $1 - 2(1+3G)p^{-\tau}$. All following arguments assume (4.14) holds. Note that $\{M_0, R_0^{(1)},\ldots,R_0^{(G)}\}$ is a feasible solution of (4.4) as $|I - \hat{\Sigma}^{(g)}(M_0 + R_0^{(g)})|_\infty = |(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})\Omega_0^{(g)}|_\infty \leq \|\Omega_0^{(g)}\|_{L_1}|\Sigma_0^{(g)} - \hat{\Sigma}^{(g)}|_\infty \leq C_M C_0(\log p/n)1/2 = \lambda_2$ and

$$\begin{aligned}
|G^{-1}\sum_{g=1}^{G}\left\{I - \hat{\Sigma}^{(g)}(M_0 + R_0^{(g)})\right\}|_\infty &\leq |G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})M_0|_\infty + |G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})R_0^{(g)}|_\infty \\
&\leq \|M_0\|_{L_1}|G^{-1}\sum_{g=1}^{G}(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})|_\infty + G^{-1}\sum_{g=1}^{G}\|R_0^{(g)}\|_{L_1}|\Sigma_0^{(g)} - \hat{\Sigma}^{(g)}|_\infty \\
&\leq C_M C_0\left\{\log p/(nG)\right\}^{1/2} + C_R C_0\left\{\log p/(nG)\right\}^{1/2} = \lambda_1.
\end{aligned}$$

Now, we find an upper bound of $|G(\hat{M} - M_0)e_j|_\infty = |\sum_{g=1}^{G}(\hat{\Omega}_1^{(g)} - \Omega_0^{(g)})e_j|_\infty$. In particular, we use

$$|\sum_{g=1}^{G}(\hat{\Omega}_1^{(g)} - \Omega_0^{(g)})e_j|_\infty \leq |\sum_{g=1}^{G}\Omega_0^{(g)}(\Sigma_0^{(g)} - \hat{\Sigma}^{(g)})\hat{\Omega}_1^{(g)}e_j|_\infty + |\sum_{g=1}^{G}\Omega_0^{(g)}(\hat{\Sigma}^{(g)}\hat{\Omega}_1^{(g)} - I)e_j|_\infty. \tag{4.15}$$

First, consider the first term in the right-hand side of (4.15). We can show that

$$
\begin{aligned}
|\sum_{g=1}^{G}\Omega_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{\Omega}_1^{(g)}e_j|_\infty &\le |\sum_{g=1}^{G}M_0(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{m}_j|_\infty + |\sum_{g=1}^{G}M_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{r}_j^{(g)}|_\infty \\
&\quad + |\sum_{g=1}^{G}R_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{m}_j|_\infty + |\sum_{g=1}^{G}R_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{r}_j^{(g)}|_\infty \\
&\le \|M_0\|_{L_1}\left\{|\sum_{g=1}^{G}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})|_\infty|\hat{m}_j|_1 + \sum_{g=1}^{G}|\Sigma_0^{(g)}-\hat{\Sigma}^{(g)}|_\infty|\hat{r}_j^{(g)}|_1\right\} \\
&\quad + \sum_{g=1}^{G}|R_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})|_\infty|\hat{m}_j|_1 + \sum_{g=1}^{G}|R_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})|_\infty|\hat{r}_j^{(g)}|_1.
\end{aligned}
$$

Using the assumptions $\|R_0^{(g)}\|_{L_1} \le C_R$ and $\sum_{g=1}^{G}\|R_0^{(g)}\|_{L_1} \le G^{1/2}C_R$, we have

$$
\begin{aligned}
|\sum_{g=1}^{G}\Omega_0^{(g)}(\Sigma_0^{(g)}-\hat{\Sigma}^{(g)})\hat{\Omega}_1^{(g)}e_j|_\infty &\le C_M C_0(G\log p/n)^{1/2}|\hat{m}_j|_1 + C_M C_0(\log p/n)^{1/2}\sum_{g=1}^{G}|\hat{r}_j^{(g)}|_1 \\
&\quad + C_R C_0(G\log p/n)^{1/2}|\hat{m}_j|_1 + C_R C_0(\log p/n)^{1/2}\sum_{g=1}^{G}|\hat{r}_j^{(g)}|_1 \\
&\le C_0(C_M+C_R)(G\log p/n)^{1/2}(|\hat{m}_j|_1 + G^{-1/2}\sum_{g=1}^{G}|\hat{r}_j^{(g)}|_1) \\
&\le C_0(C_M+C_R)(G\log p/n)^{1/2}(|m_{j,0}|_1 + G^{-1/2}\sum_{g=1}^{G}|r_{j,0}^{(g)}|_1) \\
&\le C_0(C_M+C_R)^2(G\log p/n)^{1/2}. \qquad (4.16)
\end{aligned}
$$

For the second term in the right-hand side of (4.15), note that

$$
\begin{aligned}
|\sum_{g=1}^{G}\Omega_0^{(g)}(\hat{\Sigma}^{(g)}\hat{\Omega}_1^{(g)}-I)e_j|_\infty &\le |\sum_{g=1}^{G}M_0(\hat{\Sigma}^{(g)}\hat{\Omega}^{(g)}-I)e_j|_\infty + |\sum_{g=1}^{G}R_0^{(g)}(\hat{\Sigma}^{(g)}\hat{\Omega}^{(g)}-I)e_j|_\infty \\
&\le \|M_0\|_{L_1}|\sum_{g=1}^{G}(\hat{\Sigma}^{(g)}\hat{\Omega}^{(g)}-I)e_j|_\infty + \sum_{g=1}^{G}\|R_0^{(g)}\|_{L_1}|(\hat{\Sigma}^{(g)}\hat{\Omega}^{(g)}-I)e_j|_\infty \\
&\le C_M\lambda_1 + G^{1/2}C_R\lambda_2 = C_0C_M(C_M+2C_R)(G\log p/n)^{1/2}. \qquad (4.17)
\end{aligned}
$$

By (4.15), (4.16), (4.17) and the equality $|\hat{M}-M_0|_\infty = \max_j|(\hat{M}-M_0)e_j|_\infty$, we have

$$
|\hat{M}-M_0|_\infty \le C_0(2C_M^2+4C_MC_R+C_R^2)\left(\frac{\log p}{nG}\right)^{1/2}.
$$

The proof is completed.

4.8.3 Proof of Theorem 9

By Theorem 7, we see that

$$\max_{ij} \sum_{g=1}^{G} |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \leq 2GC_M\lambda_2 \leq \delta_n, \tag{4.18}$$

with probability greater than $1 - 4Gp^{-\tau}$. We show that $S_0 = \hat{S}$ when (4.18) holds. For any $(i,j,g) \notin S_0$, we have $|\hat{\omega}_{ij}^{(g)}| = |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \leq \sum_{g=1}^{G} |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \leq \delta_n$. Therefore, we see $\tilde{\omega}_{ij}^{(g)} = 0$, which implies $\hat{S} \subset S_0$. On the other hand, for any $(i,j,g) \in S_0$, we have $|\hat{\omega}_{ij}^{(g)}| \geq |\omega_{ij,0}^{(g)}| - |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \geq |\omega_{ij,0}^{(g)}| - \sum_{g=1}^{G} |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| > \delta_n$. Therefore, we see that $\tilde{\omega}_{ij}^{(g)} \neq 0$, which implies $S_0 \subset \hat{S}$. In summary, we see that $S_0 = \hat{S}$ if (4.18) holds, which implies that $\mathrm{pr}(S_0 = \hat{S}) \geq \mathrm{pr}(\max_{ij} \sum_{g=1}^{G} |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \leq \delta_n)$. As $Gp^{-\tau} \to 0$, $\mathrm{pr}(\max_{ij} \sum_{g=1}^{G} |\hat{\omega}_{ij}^{(g)} - \omega_{ij,0}^{(g)}| \leq \delta_n) \to 1$ and the proof is completed.

Bibliography

Bair, Eric and Robert Tibshirani. 2004. "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data." *PLoS Biology* 2:511–522.

Banerjee, Onureena, Laurent El Ghaoui and Alexandre d'Aspremont. 2008. "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data." *Journal of Machine Learning Research* 9:485–516.

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Breiman, Leo. 1995. "Better Subset Regression Using the Nonnegative Garrote." *Technometrics* 37:373–384.

Breiman, Leo. 1996. "Heuristics of Instability and Stabilization in Model Selection." *The Annals of Statistics* 24:2350–2383.

Breiman, Leo and Jerome H. Friedman. 1997. "Predicting Multivariate Responses in Multiple Linear Regression." *Journal of the Royal Statistical Society. Series B* 59:3–54.

Cai, Tony, Weidong Liu and Xi Luo. 2011. "A Constrained L1 Minimization Approach to Sparse Precision Matrix Estimation." *Journal of the American Statistical Association* 106:594–607.

Edwards, David. 2000. *Introduction to graphical modelling*. Springer.

Efron, Bradley, Trevor Hastie, Iain Johnstone and Robert Tibshirani. 2004. "Least Angle Regression." *The Annals of Statistics* 32:407–499.

Elsawa, Sherine F., Luciana L. Almada, Steven C. Ziesmer, Anne J. Novak, Thomas E. Witzig, Stephen M. Ansell and Martin E. Fernandez-Zapico. 2011. "GLI2 Transcription Factor Mediates Cytokine Cross-talk in the Tumor Microenvironment." *The Journal of Biological Chemistry* 286:21524–21534.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96:1348–1360.

Fan, Jianqing, Yang Feng and Yichao Wu. 2009. "Network Exploration via the Adaptive Lasso and SCAD Penalties." *The Annals of Applied Statistics* 3:521541.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9:432–441.

Fu, Wenjiang J. 1998. "Penalized Regression: The Bridge versus the Lasso." *Journal of Computational and Graphical Statistics* 7:397–416.

Guo, Jian, Elizabeth Levina, George Michailidis and Ji Zhu. 2011. "Joint estimation of multiple graphical models." *Biometrika* 98:1–15.

Huang, Jianhua Z., Naiping Liu, Mohsen Pourahmadi and Linxu Liu. 2006. "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika* 93:85–98.

Knight, Keith and Wenjiang Fu. 2000. "Asymptotics for lasso-type estimators." *The Annals of Statistics* 28:1356–1378.

Lam, Clifford and Jianqing Fan. 2009. "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation." *The Annals of Statistics* 37:42544278.

Lee, Wonyul, Ying Du, Wei Sun, David Neil Hayes and Yufeng Liu. 2012. "Multiple Response Regression for Gaussian Mixture Models with Known Labels." *Statistical Analysis and Data Mining* 5:493–508.

Lee, Wonyul and Yufeng Liu. 2012. "Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood." *Journal of Multivariate Analysis* 111:241–255.

Lee, Wonyul and Yufeng Liu. 2013. "Joint estimation of multiple precision matrices with common structures." *Submitted to Journal of Machine Learning Research* .

Liu, Han, John Lafferty and Larry Wasserman. 2009. "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs." *Journal of Machine Learning Research* 10:2295–2328.

Meier, Lukas, Sara van de Geer and Peter Buhlmann. 2008. "The group lasso for logistic regression." *Journal of the Royal Statistical Society. Series B* 70:53–71.

Meinshausen, Nicolai and Peter Buhlmann. 2006. "High-dimensional graphs and variable selection with the Lasso." *The Annals of Statistics* 34:1436–1462.

Peng, Jie, Pei Wang, Nengfeng Zhou and Ji Zhu. 2009. "Partial Correlation Estimation by Joint Sparse Regression Models." *Journal of the American Statistical Association* 104:735–746.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi. 2002. "Hierarchical orgarnization of modularity in metabolic networks." *Science* 297:1151–1155.

Rothman, Adam J., Elizaveta Levina and Ji Zhu. 2010. "Sparse Multiple Regression with Covariance Estimation." *Journal of Computational and Graphical Statistics* 19:947–962.

Rothman, Adam J., Peter J. Bickel, Elizaveta Levina and Ji Zhu. 2008. "Sparse permutation invariant covariance estimation." *Electronic Journal of Statistics* 2:494–515.

Seike, Masahiro, Akiteru Goto, Tetsuya Okano, Elise D. Bowman, Aaron J. Schetter, Izumi Horikawa, Ewy A. Mathe, Jin Jen, Ping Yang, Haruhiko Sugimura, Akihiko Gemma, Shoji Kudoh, Carlo M. Croce and Curtis C. Harris. 2009. "MiR-21 is an EGFR-regulated anti-apoptotic factor in lung cancer in never-smokers." *PNAS* 106:12085–12090.

TCGA. 2008. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature* 455:1061–1068.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society. Series B* 58:267–288.

Turlach, Berwin A., William N. Venables and Stephen J. Wright. 2005. "Simulatneous Variable Selection." *Technometrics* 47:349–363.

Uchida, Hiroyuki, Kazunori Arita, Shunji Yunoue, Hajime Yonezawa, Yoshinari Shinsato, Hiroto Kawano, Hirofumi Hirano, Ryosuke Hanaya and Hiroshi Tokimura. 2011. "Role of sonic hedgehog signaling in migration of cell lines established from CD133-positive malignant glioma cells." *J Neurooncol* 104:697–704.

Verhaak, Roel G.W., Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael OKelly, Pablo Tamayo, Barbara A. Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, D. Neil Hayes and TCGA. 2010. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer Cell* 17:98–110.

Yoon, Hye-Young, Koichi Miura, E. Jebb Cuthbert, Kathryn Kay Davis, Bijan Ahvazi, James E. Casanova and Paul A. Randazzo. 2006. "ARAP2 effects on the actin cytoskeleton are dependent on Arf6-specific GTPase-activating-protein activity and binding to RhoA-GTP." *Journal of Cell Science* 119:4650–4666.

Yuan, Ming. 2010. "High Dimensional Inverse Covariance Matrix Estimation via Linear Programming." *Journal of Machine Learning Research* 11:2261–2286.

Yuan, Ming, Ali Ekici, Zhaosong Lu and Renato Monteiro. 2007. "Dimension reduction and coefficient estimation in multivariate linear regression." *Journal of the Royal Statistical Society. Series B* 69:329–346.

Yuan, Ming and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society. Series B* 68:49–67.

Yuan, Ming and Yi Lin. 2007. "Model selection and estimation in the Gaussian graphical model." *Biometrika* 94:19–35.

Zhang, Bin and Steve Horvath. 2005. "A General Framework for Weighted Gene Co-Expression Network Analysis." *Statistical Applications in Genetics and Molecular Biology* 4:1–45.

Zhang, Hao Helen, Yufeng Liu, Yichao Wu and Ji Zhu. 2008. "Variable selection for the multicategory SVM via adaptive sup-norm regularization." *Electronic Journal of Statistics* 2:149–167.

Zhao, Peng, Guilherme Rocha and Bin Yu. 2009. "Grouped and Hierarchical Model Selection through Composite Absolute Penalties." *The Annals of Statistics* 37:3468–3497.

Zhou, Nengfeng and Ji Zhu. 2010. "Group variable selection via a hierarchical lasso and its oracle property." *Statistics and Its Interface* 3:557–574.

Zhou, Xuan, Yu Ren, Lynette Moore, Mei Mei, Yongping You, Peng Xu, Baoli Wang, Guangxiu Wang, Zhifan Jia, Peiyu Pu, Wei Zhang and Chunsheng Kang. 2010. "Downregulation of miR-21 inhibits EGFR pathway and suppresses the growth of human glioblastoma cells independent of PTEN status." *Laboratory investigation* 90:144–155.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101:1418–1429.

Zou, Hui and Runze Li. 2008. "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models." *The Annals of Statistics* 36:1509–1533.