Published in final edited form as:
J Multivar Anal. 2012 October 1; 111: 241–255. doi:10.1016/j.jmva.2012.03.013.

Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood

Wonyul Lee and **Yufeng Liu***
Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Wonyul Lee: wonyull@email.unc.edu; Yufeng Liu: yfliu@email.unc.edu

Abstract

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. In this paper, we propose three new methods for utilizing joint information among response variables. All methods are in a penalized likelihood framework with weighted L_1 regularization. The proposed methods provide sparse estimators of conditional inverse co-variance matrix of response vector given explanatory variables as well as sparse estimators of regression parameters. Our first approach is to estimate the regression coefficients with plug-in estimated inverse covariance matrices, and our second approach is to estimate the inverse covariance matrix with plug-in estimated regression parameters. Our third approach is to estimate both simultaneously. Asymptotic properties of these methods are explored. Our numerical examples demonstrate that the proposed methods perform competitively in terms of prediction, variable selection, as well as inverse covariance matrix estimation.

Keywords

GLASSO; Inverse covariance matrix estimation; Joint estimation; LASSO; Multiple response; Sparsity

1. Introduction

Parameter estimation and variable selection are two important goals in linear regression analysis. In traditional statistical procedures, these two objectives are often achieved separately. For example, parameter estimation can be done by the least squares regression method and variable selection can be achieved by certain subset selection techniques. However, with a large number of predictors available in practice, these methods may not be

*Address for correspondence: Yufeng Liu, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, CB3260, University of North Carolina, Chapel Hill, NC 27599. yfliu@email.unc.edu. Phone: (919) 923 - 7898. Fax: (919) 962-1279.

feasible. When the dimension gets large, the least squares method may have an overfitting problem which reduces predictive accuracy. When the dimension is larger than the sample size, the least squares regression solution cannot even be calculated directly. In terms of variable selection, the all subset selection method can be unstable because the procedure is not continuous [3], and it can be computationally infeasible when the dimension is large. To solve these problems, a large number of methods have been proposed based on the regularization framework. Some well-known methods include the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani [18], the nonnegative garrote proposed by Breiman [2], and the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li [6]. These regularized methods can help to avoid overfitting. More importantly, these techniques can perform parameter estimation and variable selection simultaneously.

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 including the above methods can be applied to each response. However, this approach may not be optimal since they do not utilize the information among response variables. To solve this multi-response regression problem, Breiman and Friedman [4] proposed a method, called the curd and whey 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. [21] proposed a method based on dimension reduction. 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 et al. [19] proposed a penalized method using the max-L_1 penalty to select a common subset of explanatory variables for multiple response regression. Their method aims to select a subset which can be used as predictors for all response variables. However, this assumption may be too strong when each response has different sets of explanatory variables.

Recently, Rothman et al. [16] proposed a penalized log-likelihood approach with the multivariate Gaussian assumption. In this paper, we further extend their method and 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 variables $\mathbf{C} = \mathbf{\Sigma}^{-1}$. 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, the first proposed method maximizes a sparse penalized log-likelihood using a previously estimated inverse covariance matrix $\hat{\mathbf{C}}$. 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 $\mathbf{C}$ and the weighted L_1 penalty on the regression parameter matrix $\mathbf{B}$. Note that the joint approach is more general than that of Rothman et al. [16], which used unweighted L_1 penalty terms. Our framework allows flexible weights on the penalty terms and it is more general. To handle the computational difficulty of high dimensional problems, we recommend 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, we introduce our proposed methodology. Section 3 explores the corresponding theoretical properties. Section 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. Section 5 provides some brief results of our numerical examples. We conclude the paper with some discussion in Section 6. The proofs of the theorems are provided in Appendix.

2. Methodology

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

$$\boldsymbol{Y} = \mathbf{XB} + \mathbf{e}, \quad \text{with} \quad \mathbf{e} = [\varepsilon_1, \ldots, \varepsilon_n]^T,$$

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

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 $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma}$ is diagonal, this separate modeling approach can work well. However, when $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ in the estimation procedure of $\mathbf{B}$. Denote $\boldsymbol{\Sigma}^{-1}$ by $\mathbf{C}$. If we assume that $\boldsymbol{\Sigma}$ is known, the log-likelihood for $\mathbf{B}$ conditional on $\mathbf{X}$ is

$$-\frac{1}{2}\mathrm{tr}\{(\boldsymbol{Y} - \mathbf{XB})\mathbf{C}(\boldsymbol{Y} - \mathbf{XB})^T\}, \tag{1}$$

up to a constant not depending on $\mathbf{B}$. Interestingly, although the maximum likelihood function involves $\boldsymbol{\Sigma}$, 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 (1) does not take any advantage from the known information on $\boldsymbol{\Sigma}$. However, when we impose penalties on the likelihood, the joint method can bring some advantage in estimation.

In this paper, 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 $\boldsymbol{\Sigma}$ is involved in the joint estimation and it is often unknown, we consider three different approaches: two plug-in methods and the doubly penalized approach. The plug-in approach in Section 2.1 uses some estimator $\hat{\mathbf{C}}$ for $\mathbf{C}$ to plug in the penalized likelihood function and then estimate $\mathbf{B}$ jointly. The plug-in approach in Section 2.2 estimates $\mathbf{C}$ after plugging in a reasonable estimator of $\mathbf{B}$. The doubly

penalized approach in Section 2.3 estimates **C** and **B** simultaneously via regularizing the estimation of both **C** and **B**.

For discussion, we first assume that **Σ** is known. To regress **Y** on **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 $\widehat{\mathbf{B}}_S = (\widehat{\beta}^S_{11}, \widehat{\beta}^S_{12})$ be the least squares solution and assume that both $\widehat{\beta}^S_{11}$ and $\widehat{\beta}^S_{12}$ are positive and $\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. With the penalty parameter λ, the separate LASSO solution is given by

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

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

$$\underset{\mathbf{B}}{\operatorname{argmin}} [\operatorname{tr}\{(\boldsymbol{Y} - \mathbf{XB})\mathbf{C}(\boldsymbol{Y} - \mathbf{XB})^T\} + \lambda|\beta_{11}| + \lambda|\beta_{12}|]. \tag{3}$$

We can show that (3) is equivalent to

$$\underset{\mathbf{B}}{\operatorname{argmin}} \left[(\mathbf{B} - \widehat{\mathbf{B}}_S)\mathbf{C}(\mathbf{B} - \widehat{\mathbf{B}}_S)^T + \lambda|\beta_{11}| + \lambda|\beta_{12}| \right] \tag{4}$$

and the solution of (4) is given by

$$\widehat{\beta}_{1m} = [\widehat{\beta}^S_{1m} - \frac{\lambda}{2}(1+\rho)]_+; \quad m = 1, 2. \tag{5}$$

Compared with the separate LASSO solution (2), the solution (5) obtains more shrinkage if ρ is positive, while negative ρ results in less shrinkage. Figure 1 provides some insight on the reason why the amount of shrinkage changes with ρ for the joint method. Solid curves in Figure 1 are contour curves of $(\mathbf{B} - \hat{\mathbf{B}}_S)\mathbf{C}(\mathbf{B} - \hat{\mathbf{B}}_S)^T$ as the quadratic function of **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 (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 increases faster along the 45° line to the horizontal axis. Thus, the solution of (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 **C**.

We propose three approaches, including two plug-in methods and one joint method. In Sections 2.1 and 2.2, we introduce two different plug-in penalized likelihood methods, one is for multiple response regression and the other one is for inverse covariance estimation. In the plug-in method for multiple response regression, we estimate **C** prior to the step of regression and then use the estimator of **C** to produce a better estimator of **B**. In the plug-in

method for inverse covariance estimation, we estimate **B** first and then estimate **C** with the estimator $\hat{\mathbf{B}}$ available. In Section 2.3, we estimate **B** and **C** together via double penalization. Section 2.4 provides some guidance on three proposed methods and model selection.

2.1. Plug-in Joint Weighted LASSO Estimator

To ensure that estimation of **B** includes the information on **Σ**, 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

$$\mathrm{tr}\{(\boldsymbol{Y}-\mathbf{XB})\mathbf{C}(\boldsymbol{Y}-\mathbf{XB})^T\}+\lambda_1\sum_{j,k} w_{jk}|\beta_{jk}|. \tag{6}$$

Here λ_1 is a tuning parameter and $w_{jk} \geq 0$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, are prespecified weights for the L_1-penalty of β_{jk}. If **C** is an $m \times m$ diagonal matrix with diagonal entries $(\sigma_1^2, \ldots, \sigma_m^2)$, then $\mathbf{y}^1, \ldots, \mathbf{y}^m$ are mutually independent. In that case, the minimizer of (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 $(k=1, \ldots, m)$. However, if **C** is not diagonal, the minimizer of (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, **C** is often not available. Thus, we need to estimate it. To estimate **C**, 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 $\mathbf{N}(\mu, \boldsymbol{\Sigma}_{\mathbf{y},\mathbf{x}})$, where $\sum_{\mathbf{y},\mathbf{x}}=\begin{pmatrix} \sum_{y,y} & \sum_{y,x} \\ \sum_{x,y} & \sum_{x,x} \end{pmatrix}$. Because **Σ** is the covariance matrix of $\mathbf{y}_i$ conditioned on $\mathbf{x}_i$, it can be expressed by $\sum=\sum_{y,y}-\sum_{y,x}\sum_{x,x}^{-1}\sum_{x,y}$. Therefore, we can estimate **Σ** by first estimating $\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{x}}$. To estimate $\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{x}}$, we adapt the Graphical LASSO (GLASSO) method proposed by Friedman et al. [7]. The GLASSO method considers the problem of estimating the inverse covariance matrix in the context of sparse Gaussian graphical models [12]. This technique was also considered by Yuan and Lin [23], Banerjee et al. [1] and Rothman et al. [15].

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

$$-\mathrm{logdet}(\sum\nolimits_{\mathbf{y},\mathbf{x}}^{-1})+\frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i-\bar{\mathbf{z}})^T\sum\nolimits_{\mathbf{y},\mathbf{x}}^{-1}(\mathbf{z}_i-\bar{\mathbf{z}})+\lambda_0\left|\left|\sum\nolimits_{\mathbf{y},\mathbf{x}}^{-1}\right|\right|. \tag{7}$$

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

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

$$\underset{\mathbf{B}}{\operatorname{argmin}} \left[\operatorname{tr}\left\{(\boldsymbol{Y}-\mathbf{XB})\widehat{\mathbf{C}}(\boldsymbol{Y}-\mathbf{XB})^T\right\} + \lambda_1 \sum_{j,k} w_{jk}|\beta_{jk}| \right], \tag{8}$$

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

2.2. Plug-in Weighted Graphical LASSO Estimator

In Section 2.1, we propose a plug-in method, PWL, which estimates **C** first and then estimates **B** given $\hat{\mathbf{C}}$. In this section, we propose another plug-in method to estimate **C**. In particular, we first estimate **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

$$\underset{\mathbf{C}}{\operatorname{argmin}} \left[-n\operatorname{logdet}(\mathbf{C}) + \operatorname{tr}\left\{(\boldsymbol{Y}-\mathbf{X}\widehat{\mathbf{B}})\mathbf{C}(\boldsymbol{Y}-\mathbf{X}\widehat{\mathbf{B}})^T\right\} + \lambda_2 \sum_{s\neq t} v_{st}|c_{st}| \right], \tag{9}$$

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

2.3. Doubly Penalized Maximum Likelihood Estimator

In Sections 2.1 and 2.2, we propose two plug-in methods. PWL estimates **C** first and then estimates **B** given $\hat{\mathbf{C}}$ while PWGL estimates **B** first and then estimates **C** given $\hat{\mathbf{B}}$. In this section, we propose to estimate (**B**, **C**) simultaneously. Since $\mathbf{y}_i|\mathbf{x}_i \sim \mathbf{N}(\mathbf{B}^T\mathbf{x}_i, \boldsymbol{\Sigma})$, the log-likelihood of (**B**, **C**) conditional on **X** is

$$\frac{n}{2}\operatorname{logdet}(\mathbf{C}) - \frac{1}{2}\operatorname{tr}\{(\boldsymbol{Y}-\mathbf{XB})\mathbf{C}(\boldsymbol{Y}-\mathbf{XB})^T\}. \tag{10}$$

It can be shown that the maximum likelihood estimator of **B** is also given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Interestingly, the resulting estimator of **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 $\mathbf{y}^1, \ldots, \mathbf{y}^m$. To incorporate the information among different response variables in estimation of **B**, we propose a joint penalized method, the doubly penalized maximum likelihood (DML) estimator, by solving

$$\underset{\mathbf{B},\mathbf{C}}{\operatorname{argmin}} \left[-n\operatorname{logdet}(\mathbf{C}) + \operatorname{tr}\{(\boldsymbol{Y}-\mathbf{XB})\mathbf{C}(\boldsymbol{Y}-\mathbf{XB})^T\} + \lambda_1 \sum_{j,k} w_{jk}|\beta_{jk}| + \lambda_2 \sum_{s\neq t} v_{st}|c_{st}| \right] \tag{11}$$

Note that the objective function in (11) is not convex with respect to (**B**, **C**). The corresponding optimization can be unstable sometimes when $p \geq n$. This is because the first term in (11) can dominate the other terms if some diagonal elements of tr $(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\}$ are zeros, which may occur when $p \geq n$. This can be shown by taking a diagonal matrix **C** and increasing the values of its diagonal elements corresponding to the zero diagonal entries in tr $(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})$. As a result, the numerical solution of **C** in (11) can have some large diagonal entries. In practice, the solution of **C** with very large diagonal entries is not desirable as it leads to very small residual variances of the corresponding

response variables. We recommend to first use the plug-in method in Section 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.

2.4. Model Selection

Two plug-in methods are preferable if one of **B** and **C** is of main interest and the other is already well estimated. Another advantage of two plug-in methods is that they have lower computational cost than the joint method. On the other hand, the joint method does not require good estimate of **B** or **C**. Even though the joint method is computationally more intensive, it often performs better than two plug-in methods in the sense that it optimizes the log-likelihood of (**B**, **C**) jointly.

The tuning parameters λ_1 and λ_2 in (8), (9) and (11) control the sparsity of the resulting estimators of (**B**, **C**). 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 ($\widehat{\mathbf{B}}_{\lambda_1}^{(-k)}$, $\widehat{\mathbf{C}}_{\lambda_2}^{(-k)}$). 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,

$$\mathrm{CV}(\lambda_1)=\sum_{k=1}^{K}\left\|\boldsymbol{Y}^{(k)}-\mathbf{X}^{(k)}\widehat{\mathbf{B}}_{\lambda_1}^{(-k)}\right\|_F^2, \tag{12}$$

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,

$$\mathrm{CV}(\lambda_2)=\sum_{k=1}^{K}\left[-n_k\log\det(\widehat{\mathbf{C}}_{\lambda_2}^{(-k)})+\mathrm{tr}\left\{(\boldsymbol{Y}^{(k)}-\mathbf{X}^{(k)}\widehat{\mathbf{B}})\widehat{\mathbf{C}}_{\lambda_2}^{(-k)}(\boldsymbol{Y}^{(k)}-\mathbf{X}^{(k)}\widehat{\mathbf{B}})^T\right\}\right], \tag{13}$$

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 (12) with a prespecified λ_2 and select $\hat{\lambda}_2$ by using (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 parts, the training set and the validation set. With a pair of (λ_1, λ_2), we first estimate (**B**, **C**) 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 (12) and (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 ($\hat{\lambda}_1$, $\hat{\lambda}_2$).

3. Asymptotic Properties

To investigate a sparse regression technique, it is necessary to investigate its asymptotic behaviors. Fan and Li [6] 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 [6], Yuan and Lin [23] and Zou [25]. The technical derivation uses the results in Knight and Fu [10]. Let $\mathbf{B}^*=(\beta^*_{jk})$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, be the true regression parameter matrix and $\mathbf{C}^*=(c^*_{st})$; $s = 1, \ldots, m$, $t = 1, \ldots, m$, be the true inverse covariance matrix. Let $\mathcal{A}=\{(j, k):\beta^*_{jk} \neq 0\}$ and $\mathcal{C}=\{(s, t):c^*_{st} \neq 0\}$. 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 β^*_{jk}; $j = 1, \ldots, p$, $k = 1, \ldots, m$.
- **(A4)** The cardinality of $\mathcal{C}$, $|\mathcal{C}| = q_2 > 0$.
- **(A5)** There exists $\tilde{c}_{st}$ which is a $\sqrt{n}$-consistent estimator of c^*_{st}; $s = 1, \ldots, m$, $t = 1, \ldots, m$.

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 $\tilde{c}_{st}$s. For the theoretical analysis, we define w_{jk} and v_{st} as $w_{jk}=\frac{1}{|\tilde{\beta}_{jk}|^\gamma}$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, $\gamma > 0$, and $v_{st}=\frac{1}{|\tilde{c}_{st}|}$; $s = 1, \ldots, m$, $t = 1, \ldots, m$, respectively.

In Sections 3.1 and 3.2, we show the plug-in estimators enjoy the oracle properties. Section 3.3 develops the asymptotic theory that reveals the oracle properties of the DML solution.

3.1. Oracle properties of the PWL solution

In this section, we first show that with the known $\mathbf{C}^*$, the minimizer of (6) is consistent in variable selection and has the asymptotic normality. Then we show that with a consistent estimator of $\mathbf{C}^*$, the PWL estimator also enjoys the same properties.

Define the true regression parameter vector as $\beta^*=(\beta^*_{11}, \ldots, \beta^*_{p1}, \ldots, \beta^*_{1m}, \ldots, \beta^*_{pm})^T$. Let $\widehat{\beta^{(n)}_{\mathbf{1}}}$ be the estimator of $\boldsymbol{\beta}^*$ obtained by minimizing (6) with the penalty parameter $\lambda_{1,n}$. Let $\beta^*_{\mathcal{A}}$ be the q_1-dimensional true parameter vector which consists of nonzero components in $\boldsymbol{\beta}^*$. Let $\widehat{\beta^{(n)}_{\mathbf{1}}}_{\mathcal{A}}$ be the corresponding estimators of $\beta^*_{\mathcal{A}}$. Let $D = (\mathbf{C}^* \otimes A)_{\mathcal{A}}$ be the $q \times q$ matrix obtained by removing the $(j + (k-1)m)$-th row and column of $\mathbf{C}^* \otimes A$ for $(j, k) \notin \mathcal{A}$. Then the following lemma shows the oracle properties of the penalized likelihood estimator $\widehat{\beta^{(n)}_{\mathbf{1}}}$ with the known $\mathbf{C}^*$, as the minimizer of (6) defined previously.

Lemma 1—(Oracle properties of the minimizer of (6), $\widehat{\beta^{(n)}_{\mathbf{1}}}$, with the known $\mathbf{C}^*$) 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)* $\lim_n P(\widehat{\beta^{\mathbf{1}}}_{jk}^{(n)}=0)=1$ if $\beta_{jk}^*=0$;
2. *(Asymptotic normality)* $\sqrt{n}(\widehat{\beta^{\mathbf{1}}}_{\mathscr{A}}^{(n)}-\beta_{\mathscr{A}}^*)\rightarrow_d \mathbf{N}(0, D^{-1})$.

Lemma 1 tells us that the penalized maximum likelihood estimator with the known **C*** satisfies the oracle properties. Since **C*** is typically unknown in practice, one often uses an estimator for **C***. 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 $\widehat{\beta^{\mathbf{2}}}^{(n)}$. Let $\widehat{\beta^{\mathbf{2}}}_{\mathscr{A}}^{(n)}$ be the corresponding estimator of $\beta_{\mathscr{A}}^*$.

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

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

Theorem 1 states that with a consistent estimator of **C***, variable selection in the PWL is consistent and the resulting estimator still enjoys the asymptotic normality.

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

$$\underset{\mathbf{C}}{\operatorname{argmin}} \left[-n\log\det(\mathbf{C})+\operatorname{tr}\{(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)\mathbf{C}(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)^T\}+\lambda_2\sum_{j\neq k} v_{jk}|c_{jk}| \right], \tag{14}$$

with the known **B***. Then we show that with a consistent estimator of **B***, the PWGL estimator still enjoys the same properties.

Denote by $\hat{\mathbf{C}}^{(1)}$ the minimizer of (14) with the known **B***. Let $\widehat{\mathbf{C}}_0^{(1)}$ be the matrix obtained from $\hat{\mathbf{C}}^{(1)}$ by replacing $\widehat{c_{jk}^{(1)}}$ with 0 if $c_{jk}^*=0$. Then the following lemma shows the oracle properties of $\hat{\mathbf{C}}^{(1)}$.

Lemma 2—(Oracle properties of the minimizer of (14), $\hat{\mathbf{C}}^{(1)}$, with known **B***) Suppose that $\lambda_{2,n}n^{-\frac{1}{2}}\rightarrow 0$ and $\lambda_{2,n}\rightarrow\infty$ as $n\rightarrow\infty$. Under the conditions **(A1), (A4)** and **(A5)**, we have the following results:

1. *(Selection consistency)* $\lim_n P(\widehat{c_{jk}^{(1)}}=0)=1$ if $c_{jk}^*=0$;
2. *(Asymptotic distribution)* $\sqrt{n}(\widehat{\mathbf{C}_0^1}-\mathbf{C}^*)\rightarrow_d \arg\min V(U)$,

   where $V(U) = tr(U\boldsymbol{\Sigma}U\boldsymbol{\Sigma}) + tr(UW)$ and W is an $m\times m$ random symmetric matrix such that $vec(W)\sim \mathbf{N}(0, \Lambda)$ in which $\operatorname{cov}(w_{ij}, w_{kl}) = \operatorname{cov}(\varepsilon_{1i}\varepsilon_{1j}, \varepsilon_{1k}\varepsilon_{1l})$. The minimum is taken over all symmetric matrices U satisfying $u_{jk}=0$ *if* $c_{jk}^*=0$.

In Lemma 2, we show that the penalized maximum likelihood estimator with the known **B*** satisfies the oracle properties. Since **B*** is typically unknown in practice, one often applies

an univariate regression technique to obtain an estimator for **B***. With slight modification of Lemma 2, we can show that the PWGL solution also enjoys the oracle properties. Denote the PWGL estimator of **C*** with the penalty parameter $\lambda_{2,n}$ as $\hat{\mathbf{C}}^{(2)}$. Let $\widehat{\mathbf{C}}_0^{(2)}$ be the matrix obtained from $\hat{\mathbf{C}}^{(2)}$ by replacing $\widehat{c}_{jk}^{(2)}$ with 0 if $c_{jk}^*=0$. 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 **B***. Under the above conditions, we have the following results:

1. *(Selection consistency)* $\lim_n P(\widehat{c}_{jk}^{(2)}=0)=1$ if $c_{jk}^*=0$;
2. *(Asymptotic distribution)* $\sqrt{n}(\widehat{\mathbf{C}}_0^2-\mathbf{C}^*)\rightarrow_d \arg\min V(U)$,

   where $V(U) = \mathrm{tr}(U\boldsymbol{\Sigma}U\boldsymbol{\Sigma}) + tr(UW)$ and W is an $m \times m$ random symmetric matrix such that $vec(W) \sim \mathbf{N}(0, \Lambda)$ in which $\mathrm{cov}(w_{ij}, w_{kl}) = \mathrm{cov}(\varepsilon_{1i}\varepsilon_{1j}, \varepsilon_{1k}\varepsilon_{1l})$. The minimum is taken over all symmetric matrices U satisfying $u_{jk} = 0$ *if* $c_{jk}^*=0$.

Theorem 2 states that with a consistent estimator of **B***, the PWGL solution satisfies the oracle properties.

3.3. Oracle properties of the DML solution

In Sections 3.1 and 3.2, we establish the oracle properties of plug-in estimators. In this section, we explore oracle properties of the DML solution in which ($\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$) 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 (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 (11) which is $\sqrt{n}$-consistent.

Lemma 3—Suppose that $\lambda_{1,n}n^{-\frac{1}{2}} \rightarrow 0$ and $\lambda_{2,n}n^{-\frac{1}{2}} \rightarrow 0$. Under the conditions **(A1)–(A5)**, there exists a local minimizer of (11) such that

$$\|(vec(\widehat{\mathbf{B}})^T, vec(\widehat{\mathbf{C}})^T)^T - (vec(\mathbf{B}^*)^T, vec(\mathbf{C}^*)^T)^T\| = O_p(1/\sqrt{n}).$$

From Lemma 3, it is clear that there exists a $\sqrt{n}$-consistent doubly penalized maximum likelihood estimator. As the DML estimator of (**B***, **C***), denote by ($\hat{\mathbf{B}}^{(n)}$, $\hat{\mathbf{C}}$) the $\sqrt{n}$-consistent local solution of (11) with the penalty parameter ($\lambda_{1,n}$, $\lambda_{2,n}$). Let $\hat{\beta}^{(n)} = \mathrm{vec}(\hat{\mathbf{B}}^{(n)})$ and let $\widehat{\beta}_{\mathscr{A}}^{(n)}$ be the corresponding estimator of $\beta_{\mathscr{A}}^*$. Let $\hat{\mathbf{C}}_0$ be the matrix obtained from $\hat{\mathbf{C}}$ by replacing $\hat{c}_{jk}$ with 0 if $c_{jk}^*=0$. 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 $\lambda_{1,n}n^{-\frac{1}{2}} \rightarrow 0$ and $\lambda_{1,n}n^{-\frac{\gamma-1}{2}} \rightarrow \infty$. In addition to that, suppose that $\lambda_{2,n}n^{-\frac{1}{2}} \rightarrow 0$ and *\lamda;2,n → ∞. Under the conditions **(A1)–(A5)**, we have the following results:

1. $\lim_n P(\widehat{\beta}_{jk}^{(n)}=0)=1$ *if* $\beta_{jk}^{*}=0$;
2. $\sqrt{n}(\widehat{\beta}_{\mathcal{A}}^{(n)}-\beta_{\mathcal{A}}^{*})\to_d \mathbf{N}(0, D^{-1})$;
3. $\lim_n P(\hat{c}_{jk}=0)=1$ *if* $c_{jk}^{*}=0$;
4. $\sqrt{n}(\widehat{\mathbf{C}}_0-\mathbf{C}^{*})\to_d \arg\min V(U)$,

   where $V(U) = tr(U\Sigma U\Sigma) + tr(UW)$ and W is a $m \times m$ random symmetric matrix such that $vec(W) \sim \mathbf{N}(0, \Lambda)$ in which $\mathrm{cov}(w_{ij}, w_{kl}) = \mathrm{cov}(\varepsilon_{1i}\varepsilon_{1j}, \varepsilon_{1k}\varepsilon_{1l})$. The minimum is taken over all symmetric matrices U satisfying $u_{jk} = 0$ *if* $c_{jk}^{*}=0$.

4. Computational Algorithm

In this section, we describe computational algorithms to solve problems (8), (9), and (11). In particular, we apply the GLASSO algorithm for (9). To solve the problems (8) and (11), we apply the coordinate-descent algorithm as described in Peng et al. [14], which can be viewed as a modification of the shooting algorithm [8]. 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 $\hat{\mathbf{C}}$ by $(\hat{c}_{ij})_{m\times m}$. Then (8) is equivalent to minimizing

$$\sum_{i=1}^{n}\sum_{k,l=1}^{m}\widehat{c}_{kl}(y_{ik}-\sum_{j=1}^{p}\beta_{jk}x_{ij})(y_{il}-\sum_{j=1}^{p}\beta_{jl}x_{ij})+\lambda_1\sum_{j,k}w_{jk}|\beta_{jk}|. \tag{15}$$

Consider (15) as a function of β_{jk} with other coefficients fixed. Then the minimizer of (15) is equivalent to

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

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 $\beta_{jk}^{(old)}$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, as the initial value for $\mathbf{B}$.

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

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

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

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

Step 3 (Iteration). Repeat Step 2 until convergence. Our stopping rule is that the change of the objective function in (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. [14]. 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 (9) in the GLASSO framework. Since (9) is equivalent to minimizing

$$-\log\det(\mathbf{C}) + \mathrm{tr}\left\{\frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\widehat{\mathbf{B}})^T(\boldsymbol{Y} - \mathbf{X}\widehat{\mathbf{B}})\mathbf{C}\right\} + \frac{\lambda_2}{n}\sum_{j\neq k} v_{jk}|c_{jk}|, \tag{16}$$

we can apply the GLASSO algorithm [7] to solve (9) by substituting the sample covariance matrix with $\frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\widehat{\mathbf{B}})^T(\boldsymbol{Y} - \mathbf{X}\widehat{\mathbf{B}})$. Therefore, the algorithm for (9) proceeds as follows:

Algorithm 2: the GLASSO Algorithm for the PWGL Method

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

Step 2 (Estimator of **C**) Given $\hat{\mathbf{B}}$, apply the GLASSO algorithm to solve (16).

Next, we combine Algorithm 1 and the GLASSO algorithm to solve problem (11) for the doubly penalized method DML in Section 2.3. The algorithm can be summarized as follows:

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

Step 1 (Initial values of **B** and **C**). Set the separate LASSO solution $\beta_{jk}^{(old)}$; $j = 1, \ldots, p$, $k = 1, \ldots, m$, as the initial value for **B** and the solution of (9), $\mathbf{C}^{(old)}$, as the initial value of **C**.

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

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

This step can be solved using the Algorithm 1.

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

$$\mathbf{C}^{(new)} = \underset{\mathbf{C}}{\mathrm{argmin}}\left[\mathrm{tr}\left\{\frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\mathbf{B}^{(new)})^T(\boldsymbol{Y} - \mathbf{X}\mathbf{B}^{(new)})\mathbf{C}\right\} - \log\det(\mathbf{C}) + \frac{\lambda_2}{n}\sum_{s\neq t} v_{st}|c_{st}|\right].$$

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 (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 **B** and **C**, 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 [22] and Meier et al. [11]. 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, $(\beta_{j1}, \ldots, \beta_{jm})$; $j = 1, \ldots, p$, can be considered as p groups. Therefore, for the prescreening step, the group lasso estimator, $\hat{\mathbf{B}}^{group}$ of **B**, is given as the minimizer of the following penalized function

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

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 $\widehat{\beta}_{j1}^{group}=\ldots=\widehat{\beta}_{jm}^{group}=0$. This prescreening step can not only speed up the computation, but also improve the prediction performance as shown in our examples.

5. Numerical Examples

In this section, our proposed methods are compared with several existing methods. The first existing method we compare is the curds and whey (CW) method proposed by Breiman and Friedman [4]. 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. Separate LASSO solutions are constructed by a modification of the LARS algorithm proposed by Efron et al. [5]. The main idea of this modified LARS algorithm was also considered by Osborne et al. [13]. In this paper, some brief summaries of the results are provided. All details about numerical examples can be found in the online supplement materials.

In simulated examples, all methods are compared in two ways, **B** estimation and **C** estimation. Performance of **B** estimation is compared in terms of prediction and variable selection. For the comparison of **C** estimation, we use the entropy criterion [9, 24] which measures the difference of two matrices. In terms of prediction, overall, the proposed DML method works the best. The PWL method also works reasonably well in all cases, although it is not as accurate as the DML estimator. In the example where the true inverse covariance matrix is not sparse, LASSO gives the worst prediction performance while the other methods show similar performance. This implies that joint approaches outperform separate approaches with the joint information. DML outperforms LASSO and PWL in terms of identification of zero coefficients. We also notice that the ratios of correctly identified zeros for PWL and DML increase as the sample size increases. This supports the selection consistency shown in Section 3. When the dimension of predictor variables is low, the DML estimator gives the best performance in **C** estimation. When the dimension of predictor variables is close to the sample size, PWGL outperforms DML. Since the DML method simultaneously estimates both **B** and **C**, with a small sample size, the **C** estimation may not be as good.

We apply our methodology to a Glioblastoma multiforme (GBM) cancer data set studied by the Cancer Genome Atlas (TCGA) Research Network [17]. As noted in [20], GBM is the most common primary form of brain tumor in adults. In terms of prediction, our method, DML, performs best even though the difference between DML and the separate LASSO is not statistically significant in view of the standard errors. In terms of the number of included genes in models, PWL and DML construct sparser models than the separate LASSO. One possible explanation is that there may be some strong positive correlations among microRNAs which are response variables. As we have discussed in the toy example of Section 2, with strong positive correlations among response variables, joint methods tend to obtain more shrinkage than the separate LASSO. To explore this further, correlations among microRNAs are examined. Some strong postive correlations among the microRNAs are detected while negative correlations are not strong. Interestingly, with much fewer number of gene expressions than the separate LASSO, PWL and DML perform competitively in terms of prediction accuracy.

6. Discussion

In this paper, we have proposed three methods for utilizing joint information among response variables in a penalized likelihood framework with weighted L_1 regularization. Our theoretical investigation shows that our proposed estimators enjoy oracle properties. Simulated examples and an application to the GBM cancer data set demonstrate that our 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 distributional 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors were supported in part by NSF grant DMS-0747575 and NIH grant 5R01CA149569-03.

References

1. Banerjee O, Ghaoui LE, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research. 2008; 9:485–516.
2. Breiman L. Better subset regression using the nonnegative garrote. Technometrics. 1995; 37:373–384.
3. Breiman L. Heuristics of instability and stabilization in model selection. The Annals of Statistics. 1996; 24:2350–2383.
4. Breiman L, Friedman JH. Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society Series B. 1997; 59:3–54.
5. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004; 32:407–499.
6. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.
7. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

8. Fu WJ. Penalized regression: The bridge versus the lasso. Journal of Computational and Graphical Statistics. 1998; 7:397–416.
9. Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika. 2006; 93:85–98.
10. Knight K, Fu W. Asymptotics for lasso-type estimators. The Annals of Statistics. 2000; 28:1356–1378.
11. Meier L, van de Geer S, Buhlmann P. The group lasso for logistic regression. Journal of the Royal Statistical Society Series B. 2008; 70:53–71.
12. Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006; 34:1436–1462.
13. Osborne MR, Presnell B, Turlach BA. A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis. 2000; 20:389–404.
14. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association. 2009; 104:735–746. [PubMed: 19881892]
15. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.
16. Rothman AJ, Levina E, Zhu J. Sparse multiple regression with covariance estimation. Journal of Computational and Graphical Statistics. 2010; 19:947–962.
17. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
18. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B. 1996; 58:267–288.
19. Turlach BA, Venables WN, Wright SJ. Simulatneous variable selection. Technometrics. 2005; 47:349–363.
20. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN. TCGA. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]
21. Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society Series B. 2007; 69:329–346.
22. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B. 2006; 68:49–67.
23. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika. 2007; 94:19–35.
24. Zhu Z, Liu Y. Estimating spatial covariance using penalised likelihood with weighted l_1 penalty. Journal of Nonparametric Statistics. 2009; 21:925–942.
25. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Appendix A. Proof of Lemma 1

Appendix A.1. Asymptotic normality

Let $\tilde{\mathbf{Y}} = ((\mathbf{y}^1)^T, \ldots, (\mathbf{y}^m)^T)^T$ be the nm-dimensional response vector and $\tilde{\boldsymbol{\varepsilon}}$ be the corresponding nm-dimensional error vector which consists of ε_{ik}; $i = 1, \ldots, n$, $k = 1, \ldots, m$. Let $\tilde{\boldsymbol{\beta}} = (\beta_{11}, \ldots, \beta_{p1}, \ldots, \beta_{1m}, \ldots, \beta_{pm})^T$ be the pm-dimensional vector and $\tilde{\mathbf{X}} = \mathbf{I}_m \otimes \mathbf{X}$. Then the minimizer of (6) is equivalent to

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

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

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

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

$$V_n(\mathbf{u})-V_n(\mathbf{0})=\frac{1}{n}\mathbf{u}^T\tilde{\mathbf{X}}^T(\mathbf{C}\otimes\mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u}-\frac{2}{\sqrt{n}}\tilde{\boldsymbol{\varepsilon}}^T(\mathbf{C}\otimes\mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u}+\lambda_{1,n}\sum_{j,k} w_{jk}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|). \quad \text{(A.1)}$$

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

- If $\beta^*_{jk}=0$, then $\lambda_{1,n}w_{jk}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|)=\frac{\lambda_{1,n}}{\sqrt{n}}w_{jk}|u_{jk}|=\lambda_{1,n}n^{\frac{\gamma-1}{2}}\frac{|u_{jk}|}{(\sqrt{n}|\tilde{\beta}_{jk}|)^\gamma}\to\infty$ as $\sqrt{n}\tilde{\beta}_{jk}=O_p(1)$.
- if $\beta^*_{jk}\neq 0$, then $\lambda_{1,n}w_{jk}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|)=\frac{\lambda_{1,n}}{\sqrt{n}}w_{jk}\sqrt{n}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|)$. Note that $\frac{\lambda_{1,n}}{\sqrt{n}}\to 0, w_{jk}\to_p\frac{1}{|\beta^*_{jk}|^\gamma}$ and $\sqrt{n}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|)\to u_{jk}\operatorname{sign}(\beta^*_{jk})$. By the Slutsky's theorem, $\lambda_{1,n}w_{jk}(\left|\beta^*_{jk}+\frac{u_{jk}}{\sqrt{n}}\right|-|\beta^*_{jk}|)\to_p 0$.

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

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

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

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

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 $\widehat{\mathbf{u}}^{(n)}_{\mathcal{A}} \to_d \mathbf{N}(0, D^{-1})$. Finally, we have that $\widehat{\mathbf{u}}^{(n)}_{\mathcal{A}} = \sqrt{n}(\widehat{\boldsymbol{\beta}^{\mathbf{1}}}^{(n)}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to_d D^{-1}\mathbf{Z}_{\mathcal{A}}$ as $n \to \infty$.

Appendix A.2. Selection consistency

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

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

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

Appendix B. Proof of Theorem 1

The proof is similar to that of Lemma 1 except we replace $\mathbf{C}$ by $\widehat{\mathbf{C}}$.

Appendix B.1. Asymptotic normality

Note that (8) is equivalent to

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

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

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

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

$$V_n^*(\mathbf{u})-V_n^*(0)=V_n(\mathbf{u})-V_n(\mathbf{0})+\frac{1}{n}\mathbf{u}^T\tilde{\mathbf{X}}^T((\widehat{\mathbf{C}}-\mathbf{C})\otimes\mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u}-\frac{2}{\sqrt{n}}\tilde{\boldsymbol{\varepsilon}}^T((\widehat{\mathbf{C}}-\mathbf{C})\otimes\mathbf{I}_n)\tilde{\mathbf{X}}\mathbf{u},$$

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

Appendix B.2. Selection consistency

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

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

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

Appendix C. Proof of Lemma 2

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

$$Q(\mathbf{C})=-n\log\det(\mathbf{C})+n\mathrm{tr}(\mathbf{C}R)+\lambda_{2,n}\sum_{j\neq k}v_{jk}|c_{jk}|. \tag{C.1}$$

Appendix C.1. Selection consistency

Using the definition of $Q(\mathbf{C})$ in (C.1), define $V_n(U)$ as

$$V_n(U)=Q(\mathbf{C}^*+\tfrac{U}{\sqrt{n}})-Q(\mathbf{C}^*)$$
$$=-n\log\det((\mathbf{C}^*+\tfrac{U}{\sqrt{n}})\mathbf{C}^{*-1})+n\mathrm{tr}(\tfrac{UR}{\sqrt{n}})+\lambda_{2,n}\sum_{j\neq k}v_{jk}(|c_{jk}^*+\tfrac{u_{jk}}{\sqrt{n}}|-|c_{jk}^*|).$$

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

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

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

$$\begin{aligned}\lambda_{2,n}\sum_{j\neq k}v_{jk}(|c^*_{jk}+\tfrac{u_{jk}}{\sqrt{n}}|-|c^*_{jk}|)&=\lambda_{2,n}\sum_{c^*_{jk}=0}\frac{|u_{jk}|}{\sqrt{n}|\tilde{c}_{jk}|}+\frac{\lambda_{2,n}}{\sqrt{n}}\sum_{c^*_{jk}\neq 0}(\frac{|u_{jk}|}{|\tilde{c}_{jk}|}\mathrm{sign}(c^*_{jk})+o(1))\\&=\lambda_{2,n}\sum_{c^*_{jk}=0}\frac{|u_{jk}|}{\sqrt{n}|\tilde{c}_{jk}|}+o_p(1).\end{aligned}$$

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

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

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 $\sqrt{n}|\tilde{c}_{jk}|=O_p(1)$. Therefore, if $c^*_{jk}=0$, $P(\hat{u}_{jk}=0)\to 1$ as $n\to\infty$. This completes the proof of the variable selection consistency.

Appendix C.2. Asymptotic distribution

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

$$V_n(U)=\mathrm{tr}(U\sum U\sum)+\mathrm{tr}[\,U\sqrt{n}(R-\sum)]+o_p(1).$$

By using the Slutsky's theorem, we have that

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

Since $V_n(U)$ and $V(U)$ are both convex and $V(U)$ has a unique minimum, argmin $V_n(U)\to_d$ argmin $V(U)$. From the fact that

argmin $V_n(U)$=argmin $Q(\mathbf{C}^*+\frac{U}{\sqrt{n}})=\sqrt{n}(\widehat{\mathbf{C}}_0^1-\mathbf{C}^*)$, argmin $V_n(U)=\sqrt{n}(\widehat{\mathbf{C}}_0^1-\mathbf{C}^*)\to_d$argmin $V(U)$. This completes the proof of the asymptotic distribution.

Appendix D. Proof of Theorem 2

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

$$Q(\mathbf{C})=-n\log\det(\mathbf{C})+n\mathrm{tr}(\mathbf{C}\widehat{R})+\lambda_{2,n}\sum_{j\neq k}v_{jk}|c_{jk}|. \tag{D.1}$$

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

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

Note that

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

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

$$n\mathrm{tr}(\frac{U(\widehat{R}-R)}{\sqrt{n}})=o_p(1). \tag{D.2}$$

The left-hand side of (D.2) can be written as

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

where we add and subtract $\mathbf{XB}$ in the first term. Since $\sqrt{n}(\widehat{\mathbf{B}}-\mathbf{B})=O_p(1)$, $\frac{(\boldsymbol{Y}-\mathbf{X}\mathbf{B})^T\mathbf{X}}{\sqrt{n}}=O_p(1)$, ($\hat{\mathbf{B}}$-$\mathbf{B}$ = $o_p(1)$ and $\frac{1}{n}\mathbf{X}^T\mathbf{X}\to A$, (D.2) holds.

Appendix E. Proof of Lemma 3

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

$$Q(\mathbf{B},\mathbf{C})=-n\log\det(\mathbf{C})+\mathrm{tr}\{\mathbf{C}(\boldsymbol{Y}-\mathbf{X}\mathbf{B})^T(\boldsymbol{Y}-\mathbf{X}\mathbf{B})\}+\lambda_{1,n}\sum_{j,k}w_{jk}|\beta_{jk}|+\lambda_{2,n}\sum_{s\neq t}v_{st}|c_{S_t}|. \tag{E.1}$$

To show the results, we use the similar idea of the proof of Theorem 1 in Fan and Li [6]. It suffices to show that for any given $\delta>0$, there exists a large constant D such that

$$P\{\sup_{\|U\|=D} Q(\mathbf{B}^*+\frac{U_1}{\sqrt{n}}, \mathbf{C}^*+\frac{U_2}{\sqrt{n}})>Q(\mathbf{B}^*, \mathbf{C}^*)\} \geq 1-\delta, \tag{E.2}$$

where $U = (\mathrm{vec}(U_1)^T, \mathrm{vec}(U_2)^T)^T$. Using the definition of $Q(\mathbf{B}, \mathbf{C})$ in (E.1), define $V_n(U)$ as

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

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

$$\begin{aligned} V_n(U) \geq & -n\log\det((\mathbf{C}^*+\tfrac{U_2}{\sqrt{n}})\mathbf{C}^{*-1})+\mathrm{tr}\{(\mathbf{C}^*+\tfrac{U_2}{\sqrt{n}})(\boldsymbol{Y}-\mathbf{X}(\mathbf{B}^*+\tfrac{U_1}{\sqrt{n}}))^T(\boldsymbol{Y}-\mathbf{X}(\mathbf{B}^*+\tfrac{U_1}{\sqrt{n}}))\} \\ & -\mathrm{tr}\{\mathbf{C}^*(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)^T(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)\}+\lambda_{1,n}\sum_{\beta_{kj}\neq 0} w_{jk}(|\beta^*_{jk}+\tfrac{u_{1jk}}{\sqrt{n}}|-|\beta^*_{jk}|) \\ & +\lambda_{2,n}\sum_{c_{st}\neq 0} v_{st}(|c^*_{st}+\tfrac{u_{2st}}{\sqrt{n}}|-|c^*_{st}|) \\ = & -n\log\det((\mathbf{C}^*+\tfrac{U_2}{\sqrt{n}})\mathbf{C}^{*-1})+\mathrm{tr}\{\tfrac{U_2}{\sqrt{n}}(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)^T(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)\} \\ & +\mathrm{tr}\{(\mathbf{C}^*+\tfrac{U_2}{\sqrt{n}})(\tfrac{\mathbf{X}U_1}{\sqrt{n}})^T(\tfrac{\mathbf{X}U_1}{\sqrt{n}})\}-2\mathrm{tr}\{(\mathbf{C}^*+\tfrac{U_2}{\sqrt{n}})(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)^T(\tfrac{\mathbf{X}U_1}{\sqrt{n}})\} \\ & +\lambda_{1,n}\sum_{\beta_{kj}\neq 0} w_{jk}(|\beta^*_{jk}+\tfrac{u_{1jk}}{\sqrt{n}}|-|\beta^*_{jk}|)+\lambda_{2,n}\sum_{c_{st}\neq 0} v_{st}(|c^*_{st}+\tfrac{u_{2st}}{\sqrt{n}}|-|c^*_{st}|). \end{aligned} \tag{E.3}$$

For the first term and the second term on the right-hand side of (E.3), it has been shown in Lemma 2 that

$$-n\log\det((\mathbf{C}^*+\frac{U_2}{\sqrt{n}})\mathbf{C}^{*-1})+\mathrm{tr}\{\frac{U_2}{\sqrt{n}}(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)^T(\boldsymbol{Y}-\mathbf{X}\mathbf{B}^*)\}=\mathrm{tr}(U_2\sum U_2\sum)+\mathrm{tr}(U_2W_n).$$

Let $U_1 = \mathrm{vec}(U_1)$. For the third term on the right-hand side of (E.3), as $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to A$, note that

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

For the fourth term on the right-hand side of (E.3), we have

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

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

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

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

$$\lambda_{1,n} \sum_{\beta_{jk}^* \neq 0} w_{jk} (|\beta_{jk}^* + \frac{u_{1jk}}{\sqrt{n}}| - |\beta_{jk}^*|) = \frac{\lambda_{1,n}}{\sqrt{n}} \sum_{\beta_{jk}^* \neq 0} (\frac{|u_{1jk}|}{|\beta_{jk}|^\gamma} \operatorname{sign}(\beta_{jk}^*) + o(1)) = o_p(1),$$
$$\lambda_{2,n} \sum_{c_{st} \neq 0} v_{st} (|c_{st}^* + \frac{u_{2st}}{\sqrt{n}}| - |c_{st}^*|) = \frac{\lambda_{2,n}}{\sqrt{n}} \sum_{c_{st}^* \neq 0} (\frac{|u_{2st}|}{|\tilde{c}_{st}|} \operatorname{sign}(c_{st}^*) + o(1)) = o_p(1).$$

Therefore,

$$V_n(U) \geq \operatorname{tr}(U_2 \sum U_2 \sum) + \operatorname{tr}(U_2 W_n) + \tilde{U}_1^T (\mathbf{C}^* \otimes A) \tilde{U}_1 + \tilde{U}_1^T Z_n + o_p(1). \quad \text{(E.4)}$$

By choosing a sufficiently large D, $V_n(U) > 0$ uniformly on $\{U : \|U\| = D\}$ with the probability greater than $1\text{-}\delta$ as $\mathbf{C}^*$ and A are positive-definite, $W_n = O_p(1)$, and $Z_n = O_p(1)$. Therefore, (E.2) holds. This completes the proof of this lemma.

Appendix F. Proof of Theorem 3

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

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

Note that $(\hat{\mathbf{B}}^{(n)}, \hat{\mathbf{C}})$ is a $\sqrt{n}$-consistent local minimizer of $Q(\mathbf{B}, \mathbf{C})$. As $\hat{\mathbf{B}}^{(n)} = \operatorname{argmin}_{\mathbf{B}} Q(\mathbf{B}, \hat{\mathbf{C}})$ and $\hat{\mathbf{C}}$ is $\sqrt{n}$-consistent, the oracle properties of $\hat{\mathbf{B}}^{(n)}$ hold by Theorem 1. Similarly, since $\hat{\mathbf{C}} = \operatorname{argmin}_{\mathbf{C}} Q(\hat{\mathbf{B}}^{(n)}, \mathbf{C})$ and $\hat{\mathbf{B}}^{(n)}$ is $\sqrt{n}$-consistent, the oracle properties of $\hat{\mathbf{C}}$ hold by Theorem 2. These complete the proof of this theorem.

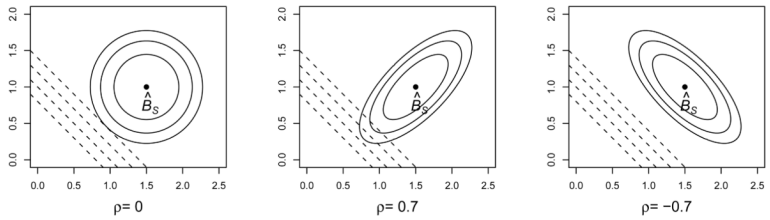

Figure 1.
Contour plots for the toy example to illustrate the change of shrinkage with ρ for the joint method.