Statistica Sinica **25** (2015), 831-851
doi:http://dx.doi.org/10.5705/ss.2013.192

JOINT ESTIMATION OF SPARSE MULTIVARIATE REGRESSION AND CONDITIONAL GRAPHICAL MODELS

Junhui Wang

University of Illinois at Chicago and City University of Hong Kong

Abstract: Multivariate regression model is a natural generalization of the classical univariate regression model for fitting multiple responses. In this paper, we propose a high-dimensional multivariate conditional regression model for constructing sparse estimates of the multivariate regression coefficient matrix that accounts for the dependency structure among the multiple responses. The proposed method decomposes the multivariate regression problem into a series of penalized conditional log-likelihood of each response conditional on the covariates and other responses. It allows simultaneous estimation of the sparse regression coefficient matrix and the sparse inverse covariance matrix. The asymptotic selection consistency and normality are established for the diverging dimension of the covariates and number of responses. The effectiveness of the proposed method is demonstrated in a variety of simulated examples as well as an application to the Glioblastoma multiforme cancer data.

Key words and phrases: Covariance selection, Gaussian graphical model, large p small n, multivariate regression, regularization.

1. Introduction

Multivariate regression model is a key statistical tool for analyzing dataset with multiple responses. A standard approach is to decompose the multivariate regression model and fit each response via a marginal univariate regression model. However, this approach is suboptimal in general as it ignores the dependency structure among the responses. For example, the expression profiles of many genes are strongly correlated due to the shared genetic variants or other unmeasured common regulators (Kendziorski et al. (2006)). With the dependency structure appropriately incorporated, one would expect a more efficient multivariate regression model in terms of both estimation and prediction. Furthermore, the dependency structure among the responses can be nicely interpreted in a graphical model under the multivariate Gaussian assumption (Edwards (2000)), where two Gaussian responses are independent conditional on other responses if the corresponding entry in the precision matrix (inverse covariance matrix) is zero.

To model the multivariate regression problem, Breiman and Friedman (1997) proposed the curd and whey method to improve the prediction performance by utilizing the dependency among responses. The curd fits a univariate regression model for each response against the covariates, and the whey refits each response against the fitted values from the curd; the method does not address the challenges when the data dimension is diverging. Yuan et al. (2007) and Chen and Huang (2012) proposed a high dimensional reduced-rank regression model, which assumes that all marginal regression functions reside in a common low-dimensional space; this approach focuses on dimension reduction and largely relies on the reduced-rank assumption. Turlach, Venables, and Wright (2005) imposed sparsity in the regression model through a L_∞-norm penalty of the coefficient matrix; this method can produce bias for model estimation due to the L_∞-norm penalty. The recent work by Rothman, Levina, and Zhu (2010), Yin and Li (2011), and Lee and Liu (2012) formulated the multivariate regression problem in a penalized log-likelihood framework, so that it allows joint estimation of the multivariate regression model and the conditional Gaussian graphical model; this formulation is computationally expensive and does not guarantee a global optimum.

We propose a penalized conditional log-likelihood formulation for the multivariate regression problem with diverging dimension. The conditional log-likelihood function is constructed for each response conditional on the covariates and other responses. The advantage here is in the inclusion of other responses in each conditional log-likelihood function, allowing joint estimation of the multivariate regression model and the dependency structure among the responses. The conditional log-likelihood function is equipped with the adaptive Lasso penalty (Zou (2006)) to facilitate the sparse estimation of the multivariate regression coefficient matrix and the precision matrix. The proposed model leads to a series of augmented adaptive Lasso regression models, that can be efficiently solved by existing optimization packages. The asymptotic properties established are estimation consistency and selection consistency with diverging dimension. The dimension of covariates and the number of responses are allowed to diverge in an exponential order of the sample size. Simulation and a data example support the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to the multivariate regression model, with an emphasis on the penalized log-likelihood method. Section 3 describes the proposed penalized conditional log-likelihood method in detail. Section 4 establishes the asymptotic selection consistency and normality for the proposed method. Simulation and an application to real data are in Section 5. Section 6 contains a discussion, and the Appendix is devoted to proofs.

2. Preliminaries

In a multivariate regression setting, supposed that the training dataset consists of $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^T \in \mathbb{R}^q$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ be the $n \times p$ design matrix and $n \times q$ response matrix, where p and q are diverging with n. Let $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})^T$ and $\mathbf{y}^k = (y_{1k}, \ldots, y_{nk})^T$ be the jth covariate and the kth response. For simplicity, the covariates and responses are centered, so that

$$\sum_{i=1}^n x_{ij} = 0, \ \sum_{i=1}^n y_{ik} = 0; \ j = 1, \ldots, p; \ k = 1, \ldots, q.$$

The multivariate linear regression model is

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{B} = (\beta_1, \ldots, \beta_q)$, with $\beta_k = (\beta_{1k}, \ldots, \beta_{pk})^T \in \mathbb{R}^p$ the regression coefficient for the kth response, and $\mathbf{e} = (e_1, \ldots, e_n)^T$, with $e_i = (e_{i1}, \ldots, e_{iq})^T \in \mathbb{R}^q$ the ith error vector. We consider a fixed design $\mathbf{X}$, so the randomness comes from the error vectors, assumed to be independent and identically sampled from a q-dimensional Gaussian distribution $N_q(0, \mathbf{\Sigma})$ with positive definite $\mathbf{\Sigma} = (\sigma_{st})_{s,t=1}^q$.

The maximum likelihood formulation of (2.1), after dropping constant terms, is

$$\min_{\mathbf{B}, \mathbf{\Omega}} \ -\log|\,\mathbf{\Omega}\,| + \operatorname{tr}\Big((\mathbf{Y} - \mathbf{X}\,\mathbf{B})\,\mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\,\mathbf{B})^T\Big), \tag{2.2}$$

with $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{st})_{s,t=1}^q$ the precision matrix. The precision matrix is closely connected with Gaussian graphical models (Edwards (2000)) since the conditional dependency structure among the responses can be fully determined by $\mathbf{\Omega}$. Specifically, $\omega_{st} = 0$ implies that the sth and tth responses are conditionally independent given the covariates and other responses.

When the dimension of covariates is large, it is generally believed that the responses only rely on a small number of covariates, while others are noise that provides no information about the responses. In addition, when the number of responses is large, the dependency structure among responses is thought sparse as some responses have little to do with each other. The penalized log-likelihood approach has been widely employed to encourage sparsity in the multivariate regression model, see Rothman, Levina, and Zhu (2010), Yin and Li (2011), and Lee and Liu (2012). The penalized likelihood approach is formulated as

$$\min_{\mathbf{B}, \mathbf{\Omega}} \ -\log|\,\mathbf{\Omega}\,| + \operatorname{tr}\Big((\mathbf{Y} - \mathbf{X}\,\mathbf{B})\,\mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\,\mathbf{B})^T\Big) + \lambda_{1n} p_1(\mathbf{B}) + \lambda_{2n} p_2(\mathbf{\Omega}), \tag{2.3}$$

where $p_1(\mathbf{B})$ and $p_2(\mathbf{\Omega})$ are sparsity-encouraging penalties, for example the adaptive Lasso penalties $p_1(\mathbf{B}) = \sum_{j,k} u_{jk}|\beta_{jk}|$ and $p_2(\mathbf{\Omega}) = \sum_{s \neq t} v_{st}|\omega_{st}|$ with

weights u_{jk} and v_{st}, and λ_{1n} and λ_{2n} are tuning parameters. To optimize (2.3), an alternating update scheme is used, that updates $\mathbf{B}$ and $\mathbf{\Omega}$, pretending the other party is fixed. When $\mathbf{B}$ is fixed, (2.3) can be solved via the graphical Lasso algorithm (Friedman, Hastie, and Tibshirani (2008)); when $\mathbf{\Omega}$ is fixed, (2.3) can be solved via the coordinate descent algorithm (Lee and Liu (2012)). However, as pointed out in Yin and Li (2011) and Lee and Liu (2012), the alternating update scheme can not guarantee the global optimum, and is often computationally expensive.

3. Proposed Methodology

A new penalized conditional log-likelihood function is developed for jointly estimating the sparse multivariate regression coefficient matrix and the sparse precision matrix. The proposed method is based on the fact that, given the model $\mathbf{y}|\mathbf{x} \sim N_q(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$ in (2.1),

$$\mathbf{y}^k|(\mathbf{X}, \mathbf{Y}^{-k}) \sim N_n(\mathbf{X}\beta_k + (\mathbf{Y}^{-k} - \mathbf{X}\mathbf{B}_{-k})\gamma_k, \tilde{\sigma}_{kk}\mathbf{I}_n), \tag{3.1}$$

for $k = 1, \ldots, q$. Here $\mathbf{Y}^{-k}$ denotes the response matrix without $\mathbf{y}^k$, $\mathbf{B}_{-k}$ denotes the coefficient matrix without β_k, $\tilde{\sigma}_{kk} = \sigma_{kk} - \mathbf{\Sigma}_{-k,k}^T\mathbf{\Sigma}_{-k,-k}^{-1}\mathbf{\Sigma}_{-k,k}$, $\mathbf{\Sigma}_{-k,k}$ is the kth column of $\mathbf{\Sigma}$ without σ_{kk}, and $\mathbf{\Sigma}_{-k,-k}$ is the submatrix of $\mathbf{\Sigma}$ without the kth row and kth column. Most importantly, β_k stays the same as in (2.1), and

$$\gamma_k = \mathbf{\Sigma}_{-k,-k}^{-1}\mathbf{\Sigma}_{-k,k} = -\frac{\mathbf{\Omega}_{-k,k}}{\omega_{kk}}, \tag{3.2}$$

where $\mathbf{\Omega}_{-k,k}$ is the kth column of $\mathbf{\Omega}$ without ω_{kk}. Since ω_{kk} is positive, it follows from (3.2) that $-\operatorname{sgn}(\gamma_k) = \operatorname{sgn}(\mathbf{\Omega}_{-k,k})$, where $\operatorname{sgn}(\gamma_k) = (\operatorname{sign}(\gamma_{1k}), \ldots, \operatorname{sign}(\gamma_{k-1,k}), \operatorname{sign}(\gamma_{k+1,k}), \ldots, \operatorname{sign}(\gamma_{q,k}))^T$ with $\operatorname{sign}(0) = 0$ for convenience. Consequently, sparsity in $\mathbf{\Omega}$ can be determined by whether $\gamma_{sk} = 0$ or not, and sparsity in $\mathbf{B}$ can be determined by whether $\beta_{jk} = 0$ or not.

To allow joint estimation of the sparse multivariate regression coefficient matrix and the sparse precision matrix, we then formulate the model in (3.1) as a series of penalized least squared regressions of each response against the covariates and other responses. Specifically, for the kth response,

$$\min_{\beta_k, \gamma_k} \|\mathbf{y}^k - \mathbf{X}\beta_k - (\mathbf{Y}^{-k} - \mathbf{X}\,\mathbf{B}_{-k})\gamma_k\|_2^2 + \lambda_{1n}p_1(\beta_k) + \lambda_{2n}p_2(\gamma_k), \tag{3.3}$$

where $\|\cdot\|_2$ is the usual Euclidean norm, $p_1(\beta_k) = \sum_{j=1}^p u_{jk}|\beta_{jk}|$ and $p_2(\gamma_k) = \sum_{s\neq k} v_{sk}|\gamma_{sk}|$ are the adaptive Lasso penalties. When $\mathbf{B}_{-k}$ in (3.3) is replaced by an initial consistent estimate $\widehat{\mathbf{B}}_{-k}^{(0)}$, the final formulation for the proposed multivariate conditional regression model is

$$\min_{\mathbf{B},\mathbf{\Gamma}} \sum_{k=1}^q \|\mathbf{y}^k - \mathbf{X}\beta_k - (\mathbf{Y}^{-k} - \mathbf{X}\,\widehat{\mathbf{B}}_{-k}^{(0)})\gamma_k\|_2^2 + \lambda_{1n}\sum_{k=1}^q p_1(\beta_k) + \lambda_{2n}\sum_{k=1}^q p_2(\gamma_k). \tag{3.4}$$

The following computing algorithm can be employed to solve (3.4).

Algorithm 1:
Step 1. Initialize $\widehat{\mathbf{B}}^{(0)}$, u_{jk} and v_{st}.
Step 2. For $k = 1, \ldots, q$, solve (3.3) for each $\hat{\beta}_k$ and $\hat{\gamma}_k$.

As computational remarks, $\widehat{\mathbf{B}}^{(0)}$ can be initialized by the separate Lasso regression ignoring the dependency structure. The weights u_{jk} and v_{sk} are set as $|\tilde{\beta}_{jk}|^{-1}$ and $|\tilde{\gamma}_{sk}|^{-1}$ as in Zou (2006), where $\tilde{\beta}_{jk}$ and $\tilde{\gamma}_{sk}$ are consistent estimates of β_{jk} and γ_{sk}, respectively. In principle, any consistent estimates can be used as long as they yield a nice bound on the estimation errors. Here we set $\tilde{\beta}_{jk}$ and $\tilde{\gamma}_{sk}$ as the solutions of (3.4) with $p_1(\beta_k)$ and $p_2(\gamma_k)$ the Lasso penalties as in Zhou, van de Geer, and Buhlmann (2009). In practice, when $\tilde{\beta}_{jk}$ or $\tilde{\gamma}_{sk}$ is 0, the corresponding weight is set to a large number to facilitate the computation.

Since (3.3) is a convex optimization problem, its global minimum can be obtained by any available adaptive Lasso regression procedure. The coordinate descent algorithm (Friedman, Hastie, and Tibshirani (2007)) can be employed to further improve the computational efficiency in solving (6). Importantly, *Step 2* fits the adaptive Lasso regression model (6) for each k, and thus can easily be parallelized and distributed to multiple computing nodes. Thus *Algorithm 1* is scalable and can efficiently handle datasets of large size.

When identifying the sparsity in the conditional graphical model defined by $\mathbf{\Omega}$, the symmetry of $\mathbf{\Omega}$ implies that $\text{sign}(\omega_{sk}) = \text{sign}(\omega_{ks})$, and thus $\text{sign}(\gamma_{sk}) = \text{sign}(\gamma_{ks})$. Consequently, additional refinement is necessary to correct the possible inconsistency in $\text{sign}(\widehat{\gamma}_{sk})$. As in Meinshausen, N. and Bühlmann, P. (2006), one can set $\widehat{\gamma}^{\wedge}_{sk} = 0$ if $\widehat{\gamma}_{sk} = 0 \wedge \widehat{\gamma}_{ks} = 0$; a less conservative way takes $\widehat{\gamma}^{\vee}_{sk} = 0$ if $\widehat{\gamma}_{sk} = 0 \vee \widehat{\gamma}_{ks} = 0$. In our numerical experiments, the less conservative way is used and the resultant selection performance in $\widehat{\Omega}$ appears to be satisfactory.

4. Asymptotic Properties

This section states the asymptotic properties of the proposed multivariate conditional regression model for diverging p and q. Let $\mathbf{B}^* = (\beta^*_{jk})$ be the true regression coefficient matrix, $\mathbf{\Omega}^* = (\omega^*_{sk})$ be the inverse of the true covariance matrix $\mathbf{\Sigma}^* = (\sigma^*_{sk})$, and $\mathbf{\Gamma}^* = (\gamma^*_{sk})$ be defined as in (3.2) with $\mathbf{\Omega}^*$. Selection accuracy is measured by the sign agreement between $(\widehat{\mathbf{B}}, \widehat{\mathbf{\Omega}})$ and $(\mathbf{B}^*, \mathbf{\Omega}^*)$, and estimation accuracy is quantified by the asymptotic normality of $n^{1/2}(\widehat{\mathbf{B}} - \mathbf{B}^*)$.

Without loss of generality, we assume that $\sigma^*_{ss} = 1$ for all s's, and take

$$\mathbf{M} = \begin{pmatrix} n^{-1}\mathbf{X}^T\mathbf{X} & 0 \\ 0 & \mathbf{\Sigma}^* \end{pmatrix}.$$

Let $\mathcal{A}_k^\beta = \{j : \beta_{jk}^* \neq 0\}$, $\mathcal{A}^\beta = \{(j,k) : j \in \mathcal{A}_k^\beta\}$, $\mathcal{A}_k^\omega = \{s : s \neq k, \omega_{sk}^* \neq 0\} = \{s : \gamma_{sk}^* \neq 0\} = \mathcal{A}_k^\gamma$, $\mathcal{A}^\omega = \{(s,k) : s \in \mathcal{A}_k^\omega\}$, $\mathcal{A} = \mathcal{A}^\beta \cup \mathcal{A}^\omega$, and $\mathcal{A}_k = \{j : j \in \mathcal{A}_k^\beta\} \cup \{p+s : s \in \mathcal{A}_k^\omega\}$. Let $d_k^\beta = |\mathcal{A}_k^\beta|$, $d_k^\omega = |\mathcal{A}_k^\omega|$, $d_k = |\mathcal{A}_k|$, and $d = \max_k\{d_k\}$. Write $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ as the minimum and maximum eigenvalues of a matrix A, and take λ_{init} and $\lambda_{1n} = \lambda_{2n} = \lambda_n$ as the tuning parameters used in the initial Lasso regression and in (3.3), respectively. Some assumptions are needed.

(A1) There exists a positive constant a_1 such that $\max_j\{n^{-1}(\mathbf{x}^j)^T\mathbf{x}^j\} \leq a_1$ and $\Lambda_{\max}(\mathbf{\Sigma}^*) \leq a_1$. In addition, $n^{-1/2}\max_i\{\mathbf{x}_i^T\mathbf{x}_i\} \to 0$ as both n and p diverge.

(A2) For some integers $1 \leq d \leq (p+q)/2$, $m \geq d$, $m+d \leq p+q$, and a positive constant k_0,

$$\frac{1}{K(d,m,k_0,\mathbf{M})} := \min_{J_0 \subset \{1,\ldots,p+q\}, |J_0| \leq d} \left(\min_{\alpha \neq 0, \|\alpha_{J_0^c}\|_1 \leq k_0 \|\alpha_{J_0}\|_1} \left(\frac{\|\mathbf{M}^{1/2}\alpha\|_2}{\|\alpha_{J_{0m}}\|_2} \right) \right) > 0.$$

(A3) If $\zeta_{\min}^* = \min_{(j,k)\in\mathcal{A}}(|\beta_{jk}^*|, |\omega_{jk}^*|)$ and $\Lambda_{\min}(d)$ are defined as below, then

$$(n\zeta_{\min}^*\Lambda_{\min}(d))^{-1} O\big(\max\big(d\lambda_{init}(\Lambda_{\min}(d))^{1/2} K(d,d,3,\mathbf{M})^2, \lambda_n d^{1/2}(\Lambda_{\min}(d))^{-1}, n^{-1}d^{1/2}\lambda_{init}, n^{1/2}d^{1/2}(\log(p+q))^{1/2}, n^{-1}d\lambda_{init}^{-2}K(d,d,3,\mathbf{M})\big)\big) \to 0.$$

In Assumption (A1), $\max_j\{n^{-1}(\mathbf{x}^j)^T\mathbf{x}^j\} \leq a_1$ is trivial and can be achieved by normalization (Zhao and Yu (2006)); $\Lambda_{\max}(\mathbf{\Sigma}^*) \leq a_1$ is assumed to guard $\mathbf{\Omega}^*$ from degeneration (Zhou, van de Geer, and Buhlmann (2009)). The condition $n^{-1/2}\max_i\{\mathbf{x}_i^T\mathbf{x}_i\} \to 0$ is necessary so that $\mathbf{Y}^{-k} - \mathbf{X}\widehat{\mathbf{B}}_{-k}^{(0)} = \mathbf{X}(\mathbf{B}_{-k}^* - \widehat{\mathbf{B}}_{-k}^{(0)}) + \mathbf{e}^{-k}$ can be well bounded. Assumption (A2) is similar to the restricted eigenvalue assumption in Bickel, Ritov, and Tsybakov (2009) and Zhou, van de Geer, and Buhlmann (2009). It implies that for any subset $S \subset \{1,\ldots,p+q\}$ with $|S| \leq d$, we have $\Lambda_{\min}(\mathbf{M}_{SS}) \geq \Lambda_{\min}(d) > 0$, where

$$\Lambda_{\min}(d) = \min_{J_0 \subset \{1,\ldots,p+q\}, |J_0| \leq d} \left(\min_{\alpha \neq 0, \alpha_{J_0^c} = 0} \left(\frac{\|\alpha^T\mathbf{M}\alpha\|_2}{\alpha_{J_0}^T\alpha_{J_0}} \right) \right).$$

Assumption (A3) is similar to the condition in Zhao and Yu (2006) and Meinshausen (2007), and implies that the nonzero β_{jk}^* and $\gamma_{sk}^* = -\omega_{sk}^*/\omega_{kk}^*$ do not decay too fast to be dominated by the noise terms.

Theorem 1 (Selection consistency)**.** *Let* (A1)$-$(A3) *hold with* $m = d$ *and* $k_0 = 3$*, the initial* $\tilde{\beta}_{jk}$ *and* $\tilde{\gamma}_{sk}$ *set as the solutions of* (3.4) *with* $p_1(\beta_k)$ *and* $p_2(\gamma_k)$ *the Lasso penalties, and* $\lambda_{1n} = \lambda_{2n} = \lambda_n$*. Then as* n*,* p*, and* q *diverge,*

$$P(\operatorname{sgn}(\widehat{\mathbf{B}}) \neq \operatorname{sgn}(\mathbf{B}^*) \text{ or } \operatorname{sgn}(\widehat{\mathbf{\Omega}}) \neq \operatorname{sgn}(\mathbf{\Omega}^*)) \longrightarrow 0,$$

when $n^{-1/2}d\lambda_{init} \to 0$, $\min\left(n^{-3}\lambda_n^2 d\lambda_{init}^2 K(d,d,3,\mathbf{M})^4, n\Lambda_{\min}(d)(\zeta^*_{\min})^2\right)(\log(p+q))^{-1} \to \infty$, *and* $(n\Lambda_{\min}(d))^{-1}\max(\lambda_{init}d, n^{-1/2}\lambda_{init}d(\log(p+q))^{1/2}, n^{-1}\lambda_{init}^2 d) \to 0$.

Theorem 2 (Asymptotic normality)**.** *If the conditions of Theorem* 1 *hold,* $s_k^2 = \tilde{\sigma}^*_{kk}\alpha^T\mathbf{M}^{-1}_{\mathcal{A}_k,\mathcal{A}_k}\alpha$ *where* α *is any* $|\mathcal{A}_k| \times 1$ *vector with unit length, and* $\mathbf{M}_{\mathcal{A}_k,\mathcal{A}_k}$ *is the principle submatrix of* $\mathbf{M}$ *defined by* $\mathcal{A}_k$*, then*

$$n^{1/2}s_k^{-1}\alpha^T\left(\begin{pmatrix}\widehat{\beta}_k\\ \widehat{\gamma}_k\end{pmatrix} - \begin{pmatrix}\beta_k^*\\ \gamma_k^*\end{pmatrix}\right) \xrightarrow{d} N(0,1) \quad \textit{for any } k,$$

when $d\lambda_n(p+q)^{-1} \to 0$, $n^{-1/2}\lambda_n d^{1/2}(\Lambda_{\min}(d)\zeta^*_{\min})^{-1} \to 0$, *and* $n^{-1/2}\lambda_{init}$ $d(\Lambda_{\min}(d))^{-1/2} \to 0$.

Thus, with consistent initial estimates of $\mathbf{B}$ and $\mathbf{\Omega}$, the proposed multivariate conditional regression model is able to achieve both selection consistency and the asymptotic normality for diverging p and q. The condition $\lambda_{1n} = \lambda_{2n} = \lambda_n$ is assumed for ease of presentation, similar results can be obtained for different λ_{1n} and λ_{2n} with slightly modified rate conditions. Both theorems can be established for fixed p and q following Huang, Ma, and Zhang (2008b), and we omit the details.

5. Numerical Experiments

This section examines the effectiveness of the proposed multivariate conditional regression model on a variety of simulated examples, and considers an application to the Glioblastoma Cancer Dataset (TCGA (2008)). The proposed model with the adaptive Lasso penalty, denoted as aMCR, is compared against the alternative updating algorithm in (2.3) (ALT; Rothman, Levina, and Zhu (2010); Yin and Li (2011); Lee and Liu (2012)), and the separate Lasso regression (SEP; ignoring the dependency structure among y_k's).

The comparison is conducted with respect to the estimation and selection accuracy of $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Omega}}$. The estimation accuracy of $\widehat{\mathbf{B}}$ is measured by the Frobenius norm $\|\mathbf{\Delta}_B\|_F = \left(\sum_{i,j}(\mathbf{\Delta}_B)_{ij}^2\right)^{1/2}$, the matrix 1-norm $\|\mathbf{\Delta}_B\|_1 = \max_j\sum_i|(\mathbf{\Delta}_B)_{ij}|$, and the matrix ∞-norm $\|\mathbf{\Delta}_B\|_\infty = \max_i\sum_j|(\mathbf{\Delta}_B)_{ij}|$, where $\mathbf{\Delta}_B = \widehat{\mathbf{B}} - \mathbf{B}^*$. The estimating accuracy of $\widehat{\mathbf{\Omega}}$ is not reported as the primary interest is the sparsity inferred by $\widehat{\mathbf{\Omega}}$, and the proposed method does not produce $\widehat{\mathbf{\Omega}}$ directly. The selection accuracy of $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Omega}}$ is measured by the symmetric difference

$$\text{Dist}(\widehat{\mathcal{A}}^\beta, \mathcal{A}^\beta) = \frac{|\widehat{\mathcal{A}}^\beta \backslash \mathcal{A}^\beta| + |\mathcal{A}^\beta \backslash \widehat{\mathcal{A}}^\beta|}{pq};$$
$$\text{Dist}(\widehat{\mathcal{A}}^\omega, \mathcal{A}^\omega) = \frac{|\widehat{\mathcal{A}}^\omega \backslash \mathcal{A}^\omega| + |\mathcal{A}^\omega \backslash \widehat{\mathcal{A}}^\omega|}{q^2},$$

where $\widehat{\mathcal{A}}^{\beta}$ and $\widehat{\mathcal{A}}^{\omega}$ are the active sets defined by $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Omega}}$, and $|\cdot|$ denotes the set cardinality. We also report the specificity (Spe), sensitivity (Sen) and Matthews correlation coefficient (Mcc) scores, defined as

$$\text{Spe} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{Sen} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{Mcc} = \frac{\text{TP}\times\text{TN}-\text{FP}\times FN}{\sqrt{(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TP}+\text{FP})(\text{TN}+\text{FN})}},$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives in identifying the nonzero elements in $\widehat{\mathbf{B}}$ or $\widehat{\mathbf{\Omega}}$, and "positive" refers to the nonzero entries.

The tuning parameters λ_{1n} and λ_{2n} in (2.3) and (3.3) control the tradeoff between the sparsity and the estimation accuracy of the multivariate regression models. In the numerical experiments, we employed the Bayesian information criterion (BIC; Schwarz (1978)) to select tuning parameters; this is known to perform well in tuning penalized log-likelihood models (Wang, Li, and Tsai (2007)). The BIC criterion for the kth conditional regression model is

$$\text{Bic}(\lambda_{1n},\lambda_{2n}) = \|\mathbf{y}^k - \mathbf{X}\widehat{\beta}_{k,\lambda} - (\mathbf{Y}^{-k} - \mathbf{X}\,\widehat{\mathbf{B}}^{(0)}_{-k})\widehat{\gamma}_{k,\lambda}\|_2^2 + \log(n)(|\widehat{\mathcal{A}}^{\beta}_{k,\lambda}| + |\widehat{\mathcal{A}}^{\omega}_{k,\lambda}|),$$

where $\widehat{\beta}_{k,\lambda}$, $\widehat{\gamma}_{k,\lambda}$, $\widehat{\mathcal{A}}^{\beta}_{k,\lambda}$, and $\widehat{\mathcal{A}}^{\omega}_{k,\lambda}$ are estimated based on $(\lambda_{1n},\lambda_{2n})$. The selected $(\lambda_{1n,k},\lambda_{2n,k})$ are obtained by minimizing $\text{Bic}(\lambda_{1n},\lambda_{2n})$ through a grid search on a two-dimensional equally-spaced grid $(10^{-3+(s-1)/3}, 10^{-3+(t-1)/3})$; $s,t = 1,\ldots,19$. The selected $(\lambda_{1n,k},\lambda_{2n,k})$ might be different from one conditional regression model to another, which appears to yield better performance than restricting to common tuning parameters. BIC is known to yield suboptimal performance in high-dimensional setting, and various modifications have been proposed, such as extended BIC (Chen and Chen (2008)). A referee points out that the conditional likelihood in $\text{Bic}(\lambda_{1n},\lambda_{2n})$ treats $\mathbf{Y}^{-k} - \mathbf{X}\,\widehat{\mathbf{B}}^{(0)}_{-k}$ as a fixed covariate, and thus the model complexity might be more involved than the number of nonzero coefficients. As an alternative, one could use cross validation (Lee and Liu (2012)) at the expense of some computation.

5.1. Simulated examples

Two simulations are considered. The first follows the setup in Li and Gui (2006), Fan, Feng, and Wu (2009), Peng et al. (2009), and Yin and Li (2011). Each entry of the precision matrix $\mathbf{\Omega}$ was generated from the product of a Bernoulli random variable with success rate proportional to $1/q$ and a uniform random variable on $[-1,-0.5]\cup[0.5,1]$. For each row, all off-diagonal entries were divided by the sum of the absolute value of the off-diagonal entries multiplied by

3/2. The precision matrix $\mathbf{\Omega}$ was obtained by symmetrizing the generated matrix and setting the diagonal entries to 1. Each entry of the coefficient matrix $\mathbf{B}$ was generated from the product of a Bernoulli random variable with success rate proportional to $1/p$ and a uniform random variable on $[-1, -v_m] \cup [v_m, 1]$, where v_m was the minimum absolute value of the nonzero entries in $\mathbf{\Omega}$. With the generated $\mathbf{\Omega}$ and $\mathbf{B}$, each entry of the covariate matrix $\mathbf{X}$ was generated independently from Bern(1/2), and the response vector was generated as $Y|X = x \sim N_q(\mathbf{B}^T x, \mathbf{\Omega}^{-1})$. Six models were considered, and for each given model, a training sample of n observations $(\mathbf{x}_i, \mathbf{y}_i);\ i = 1, \ldots, n$ was generated.

Model 1: $(p, q, n) = (100, 100, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 3/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2/q$;
Model 2: $(p, q, n) = (50, 50, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 4/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2/q$;
Model 3: $(p, q, n) = (10, 25, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 3.5/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2/q$;
Model 4: $(p, q, n) = (200, 1000, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 20/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 1.5/q$;
Model 5: $(p, q, n) = (200, 800, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 25/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 1.5/q$;
Model 6: $(p, q, n) = (200, 400, 150)$, where $P(\mathbf{B}_{ij} \neq 0) = 20/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2.5/q$.

The second simulations were similar to the first in generating $\mathbf{\Omega}$ and $\mathbf{B}$. With the generated $\mathbf{\Omega}$ and $\mathbf{B}$, each entry of the covariate matrix $\mathbf{X}$ was generated independently standard normal, and the response vector was generated from $Y|X = x \sim N_q(\mathbf{B}^T x, \mathbf{\Omega}^{-1})$. Two models were considered, and for each given model, a training sample of n observations $(\mathbf{x}_i, \mathbf{y}_i);\ i = 1, \ldots, n$ was generated.

Model 7: $(p, q, n) = (10, 25, 250)$, where $P(\mathbf{B}_{ij} \neq 0) = 3.5/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2/q$;
Model 8: $(p, q, n) = (200, 400, 150)$, where $P(\mathbf{B}_{ij} \neq 0) = 20/p$ and $P(\mathbf{\Omega}_{ij} \neq 0) = 2.5/q$.

Each model was replicated 50 times, with the averaged performance measures and the estimated standard errors reported in Tables 1 and 2. All simulations were done using R 3.0.1 on a 8-core PC with 3.4 GHz CPU and 16G memory.

The proposed aMCR delivers superior numerical performance, in terms of estimation and selection accuracy of $\mathbf{B}$ and $\mathbf{\Omega}$, against other competitors across all six simulated examples. In Tables 1 and 2, we only report the numerical performance of ALT on models 2, 3, and 7, due to the computational burden of running ALT on other models with larger dimensions. Although the performance of ALT might be improved if some random start algorithm were employed to partially overcome the issue of local minima, the inefficient alternating algorithm becomes a major obstacle in applying ALT to analyze large-dimensional datasets. Here the running times for fitting and tuning ALT and aMCR on Model 3 with

Table 1. Averaged performance measures regarding $\widehat{\mathbf{B}}$ (with estimated standard errors in parentheses) of aMCR, ALT, and SEP, over 50 replications.

	Estimation Accuracy			Selection Accuracy			
	$\|\boldsymbol{\Delta}_B\|_F$	$\|\boldsymbol{\Delta}_B\|_1$	$\|\boldsymbol{\Delta}_B\|_\infty$	Dist	Spe	Sen	Mcc
Model 1: $(p,q,n)=(100,100,250)$							
SEP	37.7(0.41)	2.84(0.055)	2.86(0.103)	0.01(0.000)	0.99(0.000)	0.55(0.005)	0.64(0.004)
ALT	–	–	–	–	–	–	–
aMCR	18.6(0.30)	2.01(0.040)	2.19(0.055)	0.01(0.000)	1.0(0.000)	0.55(0.005)	0.68(0.004)
Model 2: $(p,q,n)=(50,50,250)$							
SEP	19.0(0.261)	2.75(0.062)	2.63(0.048)	0.03(0.001)	0.96(0.001)	0.63(0.007)	0.58(0.005)
ALT	13.2(0.764)	2.49(0.063)	2.53(0.062)	0.07(0.003)	0.86(0.006)	0.77(0.016)	0.46(0.009)
aMCR	10.3(0.260)	2.04(0.062)	2.22(0.048)	0.02(0.001)	0.99(0.000)	0.60(0.006)	0.69(0.005)
Model 3: $(p,q,n)=(10,25,250)$							
SEP	4.16(0.073)	2.64(0.053)	1.45(0.036)	0.09(0.003)	0.84(0.006)	0.76(0.008)	0.60(0.007)
ALT	3.94(0.099)	2.64(0.059)	1.48(0.041)	0.11(0.003)	0.79(0.020)	0.81(0.010)	0.59(0.014)
aMCR	3.28(0.074)	2.18(0.058)	1.31(0.039)	0.07(0.002)	0.96(0.003)	0.69(0.009)	0.71(0.007)
Model 4: $(p,q,n)=(200,1000,250)$							
SEP	447.2(1.69)	10.37(0.121)	4.32(0.250)	0.01(0.000)	1.0(0.000)	0.56(0.002)	0.63(0.001)
ALT	–	–	–	–	–	–	–
aMCR	235.3(0.96)	7.06(0.090)	3.32(0.078)	0.01(0.000)	1.0(0.000)	0.54(0.001)	0.66(0.001)
Model 5: $(p,q,n)=(200,800,250)$							
SEP	355.1(1.53)	8.83(0.104)	4.05(0.075)	0.01(0.000)	1.0(0.000)	0.55(0.001)	0.63(0.001)
ALT	–	–	–	–	–	–	–
aMCR	186.4(0.86)	6.19(0.090)	3.28(0.063)	0.01(0.000)	1.0(0.000)	0.54(0.001)	0.66(0.001)
Model 6: $(p,q,n)=(200,400,150)$							
SEP	177.6(0.96)	5.34(0.070)	4.15(0.213)	0.01(0.000)	1.0(0.000)	0.56(0.002)	0.63(0.002)
ALT	–	–	–	–	–	–	–
aMCR	93.6(0.71)	3.80(0.063)	3.01(0.057)	0.01(0.000)	1.0(0.000)	0.55(0.002)	0.66(0.001)
Model 7: $(p,q,n)=(10,25,250)$							
SEP	1.07(0.021)	1.34(0.028)	0.74(0.016)	0.09(0.002)	0.80(0.005)	0.90(0.005)	0.67(0.006)
ALT	1.07(0.035)	1.36(0.027)	0.76(0.017)	0.10(0.006)	0.72(0.024)	0.92(0.008)	0.63(0.017)
aMCR	0.68(0.017)	1.05(0.029)	0.59(0.015)	0.04(0.002)	0.93(0.003)	0.87(0.005)	0.80(0.006)
Model 8: $(p,q,n)=(200,400,150)$							
SEP	173.5(0.57)	8.96(0.081)	5.75(0.053)	0.06(0.001)	0.88(0.000)	0.87(0.001)	0.57(0.001)
ALT	–	–	–	–	–	–	–
aMCR	111.0(0.34)	6.92(0.047)	4.62(0.058)	0.04(0.000)	0.94(0.000)	0.85(0.001)	0.68(0.001)

19^2 tuning parameters were127.7 seconds and 17.8 seconds, respectively. On Model 4, the running time for aMCR with 19^2 tuning parameters was around 30 minutes, whereas we could not reach results on ALT. The computing time of aMCR can be further improved if parallelized over more computing nodes.

In Tables 1 and 2, the advantage of aMCR and ALT over SEP demonstrates that inclusion of the covariance matrix in (2.3) and (3.3) is helpful in identifying the sparsity in $\mathbf{B}$ and $\boldsymbol{\Omega}$ and thus in estimating $\mathbf{B}$. As for the selection accuracy, aMCR yields higher Spe and Mcc but lower Sen in most examples. This is due

Table 2. Averaged performance measures regarding $\widehat{\mathbf{\Omega}}$ (with estimated standard errors in parentheses) of aMCR, ALT, and SEP, over 50 replications.

	Selection Accuracy			
	Dist	Spe	Sen	Mcc
Model 1: $(p,q,n)=(100,100,250)$				
SEP	0.09(0.001)	0.82(0.001)	0.77(0.006)	0.31(0.003)
ALT	–	–	–	–
aMCR	0.02(0.000)	0.99(0.000)	0.47(0.007)	0.61(0.005)
Model 2: $(p,q,n)=(50,50,250)$				
SEP	0.09(0.001)	0.82(0.002)	0.77(0.006)	0.41(0.004)
ALT	0.05(0.009)	0.93(0.021)	0.52(0.028)	0.53(0.016)
aMCR	0.05(0.001)	0.99(0.000)	0.50(0.007)	0.64(0.005)
Model 3: $(p,q,n)=(10,25,250)$				
SEP	0.09(0.003)	0.83(0.004)	0.77(0.010)	0.53(0.007)
ALT	0.11(0.013)	0.83(0.038)	0.57(0.038)	0.47(0.023)
aMCR	0.08(0.022)	0.99(0.012)	0.54(0.009)	0.65(0.007)
Model 4: $(p,q,n)=(200,1000,250)$				
SEP	0.07(0.000)	0.86(0.000)	0.73(0.001)	0.12(0.000)
ALT	–	–	–	–
aMCR	0.03(0.000)	1.0(0.000)	0.38(0.002)	0.52(0.001)
Model 5: $(p,q,n)=(200,800,250)$				
SEP	0.08(0.000)	0.85(0.000)	0.73(0.002)	0.13(0.000)
ALT	–	–	–	–
aMCR	0.00(0.000)	1.0(0.000)	0.39(0.002)	0.53(0.002)
Model 6: $(p,q,n)=(200,400,150)$				
SEP	0.08(0.000)	0.83(0.000)	0.75(0.002)	0.17(0.000)
ALT	–	–	–	–
aMCR	0.01(0.000)	1.0(0.000)	0.41(0.003)	0.55(0.002)
Model 7: $(p,q,n)=(10,25,250)$				
SEP	0.09(0.002)	0.83(0.002)	0.77(0.009)	0.52(0.006)
ALT	0.09(0.010)	0.89(0.027)	0.52(0.037)	0.49(0.017)
aMCR	0.16(0.002)	0.90(0.002)	0.70(0.010)	0.57(0.007)
Model 8: $(p,q,n)=(200,400,150)$				
SEP	0.07(0.000)	0.86(0.000)	0.58(0.002)	0.15(0.001)
ALT	–	–	–	–
aMCR	0.08(0.000)	0.92(0.000)	0.51(0.003)	0.19(0.001)

to the fact that sparse models are preferred by the less conservative rule $\widehat{\gamma}^{\vee}$, BIC criterion, and aMCR itself when the correlations among the responses are positive (Lee and Liu (2012)). Although sparser models are produced, aMCR yields smaller symmetric difference than SEP. As for the estimation accuracy of $\mathbf{B}$, it is clear that aMCR outperforms SEP under all three metrics of $\widehat{\mathbf{B}}-\mathbf{B}^*$. This suggests that the proposed model can improve not only the accuracy of identified nonzero entries in the precision matrix, but also the accuracy of estimating the

Table 3. Numerical experiments on Model 3 for the effect of various factors on aMCR. The performance measures are averaged over 50 replications.

	Estimation Accuracy			Selection Accuracy			
	$\|\mathbf{\Delta}_B\|_F$	$\|\mathbf{\Delta}_B\|_1$	$\|\mathbf{\Delta}_B\|_\infty$	Dist	Spe	Sen	Mcc
BIC	3.28(0.074)	2.18(0.058)	1.31(0.039)	0.07(0.002)	0.96(0.003)	0.69(0.009)	0.71(0.007)
CV	3.67(0.088)	2.51(0.063)	1.51(0.036)	0.11(0.003)	0.79(0.010)	0.78(0.010)	0.56(0.010)
$\widehat{\lambda}^\vee$	3.28(0.074)	2.18(0.058)	1.31(0.039)	0.07(0.002)	0.96(0.003)	0.69(0.009)	0.71(0.007)
$\widehat{\lambda}^\wedge$	3.16(0.072)	2.20(0.058)	1.30(0.037)	0.07(0.003)	0.94(0.003)	0.72(0.009)	0.69(0.007)

Table 4. Averaged performance measures regarding $\widehat{\mathbf{\Omega}}$ of aMCR, aMCR1, ALT and SEP, over 50 replications of Model 3. Here aMCR1 is the solution of (2.3) subject to the constraints of the selected non-zero entries by aMCR.

	Estimation Accuracy		
	$\|\mathbf{\Delta}_B\|_F$	$\|\mathbf{\Delta}_B\|_1$	$\|\mathbf{\Delta}_B\|_\infty$
SEP	1.98(0.033)	1.13(0.028)	1.13(0.028)
ALT	2.05(0.066)	1.19(0.065)	1.19(0.065)
aMCR	1.75(0.028)	1.05(0.025)	1.05(0.025)
aMCR1	1.94(0.043)	1.09(0.027)	1.09(0.027)

multivariate regression coefficient matrix.

Through additional numerical experiments on Model 3, comparisons between BIC or cross validation (CV), and conservative $\widehat{\gamma}^\wedge$ or less conservative $\widehat{\gamma}^\vee$ are summarized in Table 3. When aMCR is equipped with either CV or conservative $\widehat{\gamma}^\wedge$, the sensitivity can be improved but other performance measures including the specificity might deteriorate. In practice, it is recommended that the model selection criterion and the way of setting $\widehat{\gamma}$ need to be specified based on the preference between the sensitivity and specificity.

It is of interest to examine the effect of $\widehat{\mathbf{B}}$ on estimation of $\mathbf{\Omega}$, which can be obtained by solving (2.3) after plugging in $\widehat{\mathbf{B}}$. Table 4 summarizes the estimation performance measures regarding $\widehat{\mathbf{\Omega}}$ on Model 3. It is evident that the obtained $\widehat{\mathbf{\Omega}}$ by aMCR yields the smallest estimation error, followed by aMCR1 which solves (2.3) subject to the constraints of the selected non-zero entries by aMCR.

5.2. An application

We applied the proposed multivariate conditional regression model to a Glioblastoma multiforme (GBM) cancer dataset studied by the Cancer Genome Atlas (TCGA) Research Network (TCGA (2008); Verhaak et al. (2010)). GBM is the most common and most aggressive malignant primary brain tumor in adults. The original dataset collected by TCGA consists of 202 samples, 11,861 gene expression values and 534 microRNA expression values. One primary goal of the study was to regress the microRNA expressions on the gene expressions and to

Table 5. Averaged predictive square errors, numbers of selected genes and their estimated standard errors over 50 replications.

	Pse	Num.gene
SEP	1.21(0.011)	74.9(2.22)
ALT	–	–
aMCR	1.19(0.012)	65.2(1.75)

model how the microRNAs regulate the gene expressions. It is also of interest to construct the underlying network among the microRNAs. The proposed model can achieve these goals simultaneously, in that the sparse coefficient matrix reveals the regulatory relationship among the microRNA and gene expressions, and the sparse precision matrix can be interpreted as the dependency structure among the microRNAs.

Some preliminary data cleaning was conducted to remove missing values and to prescreen the less expressed genes and microRNAs as in TCGA (2008) and Lee and Liu (2012). Thus 6 samples with missing values were removed, and 196 complete samples remained in the dataset. Further, the genes and microRNAs were sorted based on their corresponding median absolute deviation (MAD), and the top 500 genes and top 20 microRNAs with large MADs were selected.

The dataset was randomly split into a training set with 120 samples and a test set with 76 samples. On the training set, each method was fitted to estimate the multivariate regression coefficient matrix and the precision matrix. Since the truth is unknown, estimation performance was measured by the predictive square error (Pse) estimated on the test set,

$$\text{Pse} = |\text{test set}|^{-1} \sum_{\text{test set}} \|\mathbf{Y}_i - \widehat{\mathbf{Y}}_i\|_F^2.$$

The numbers of the selected genes by each method are also reported.

The averaged Pse and numbers of selected genes, as well as their estimated standard errors based on 50 replications, are reported in Table 5.

The proposed aMCR yields sparser multivariate regression model and achieves smaller Pse than the separate regression model. This agrees with the conclusion in Lee and Liu (2012), and the sparser regression model is due to the fact that the joint estimation method is able to obtain more shrinkage when strong positive correlations are present among the selected microRNAs. The numerical performance of ALT is not reported due to the computational burden, but we note that in Lee and Liu (2012), the Pse of the ALT was 1.23 (0.032) and the number of selected genes was 78.0 (32.15) based on a slightly smaller dataset.

Figure 1 displays the estimated conditional dependency structure among the microRNAs based on the estimated precision matrix of the microRNAs. Compared with the results in Lee and Liu (2012), the graphical structure in Figure

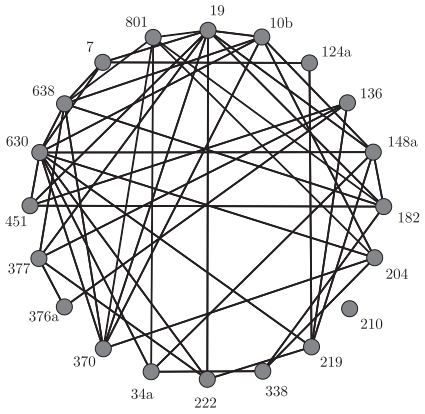

Figure 1. The dependency network of the selected microRNAs based on the estimated sparse precision matrix.

1 captures the strong positive correlations among the selected microRNA pairs, including the tuple of hsa.mir.136, hsa.mir.376a and hsa.mir.377. It produces a sparser dependency structure than that in Lee and Liu (2012), and rules out more microRNA pairs with weak correlations, such as hsa.mir.bart19 and hsa.mir.124a (with pairwise correlation -0.12).

6. Summary

This article proposes a method for jointly estimating the multivariate regression model and the dependency structure among the multiple responses. The method is formulated as a penalized conditional log-likelihood function, leading to efficient computation and superior numerical performance. Asymptotic estimation and selection consistencies are established for diverging dimensions and numbers of responses. The penalized conditional log-likelihood formulation can be extended to a general framework without the Gaussian distributional assumption of Finegold and Drton (2011) and Lee et al. (2012). In our application, as pointed out in Lee et al. (2012), there are multiple subtypes of microRNAs in the GBM dataset, and hence a Gaussian mixture distribution is more reasonable for modeling the distribution of the microRNAs. Future work along this direction is currently under investigation.

Acknowledgement

The author would like to thank Wonyul Lee and Yufeng Liu (University of North Carolina at Chapel Hill) for sharing their code on the alternative updating algorithm and the Glioblastoma multiforme cancer dataset. The author also

thank the Editor, an associate editor, and two referees for their constructive comments and suggestions, which have led to a significantly improved paper.

Appendix

Proof of Theorem 1. We establish upper bounds for $P(\mathrm{sgn}(\hat{\beta}_k) \neq \beta_k^*)$ and $P(\hat{\gamma}_k \neq \gamma_k^*)$, where $\hat{\beta}_k$ and $\hat{\gamma}_k$ are the solutions to

$$\min_{\beta_k,\gamma_k} \|\mathbf{y}^k - \mathbf{X}\beta_k - \widehat{\mathbf{y}}^{-k}\gamma_k\|^2 + \lambda_n\Big(\sum_{j=1}^{p} u_{jk}|\beta_{jk}| + \sum_{s\neq k} v_{sk}|\gamma_{sk}|\Big), \tag{A.1}$$

where $\widehat{\mathbf{y}}^{-k} = \mathbf{y}^{-k} - \mathbf{X}\,\widehat{\mathbf{B}}_{-k}^{(0)}$ is a surrogate of $\mathbf{e}^{-k} = \mathbf{y}^{-k} - \mathbf{X}\,\mathbf{B}_{-k}^*$. Based on the model assumption (3.1), we have

$$\mathbf{y}^k = \mathbf{X}\beta_k^* + \widehat{\mathbf{y}}^{-k}\gamma_k^* + \boldsymbol{\xi}_k + \boldsymbol{\epsilon}_k, \tag{A.2}$$

where $\boldsymbol{\xi}_k = (\mathbf{e}^{-k} - \widehat{\mathbf{y}}^{-k})\gamma_k^* = \mathbf{X}(\widehat{\mathbf{B}}_{-k}^{(0)} - \mathbf{B}_{-k}^*)\gamma_k^*$ and $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}_n, \tilde{\sigma}_{kk}^*\,\mathbf{I}_n)$. If $\zeta = (\beta_k^T, \gamma_k^T)^T$ is the augmented coefficient vector, and $\mathbf{Z} = (\mathbf{X}, \widehat{\mathbf{y}}^{-k})$ is the augmented covariate matrix, $r = (u_k^T, v_k^T)^T$, (A.1) can be simplified to

$$\min_{\widetilde{\beta}} \|\mathbf{y}^k - \mathbf{Z}\zeta\|^2 + \lambda_n \sum_{j=1}^{p+q-1} r_j|\zeta_j|. \tag{A.3}$$

We now verify the conditions (A.4) and (A.5) in Lemma A.1. For simplicity, let

$$\mathcal{T} = \Big\{\max_{j,s}\ n^{-1}(\mathbf{X}^j)^T\mathbf{e}^s \leq a_1^2(8n^{-1}\log(p+q))^{1/2}\Big\},$$

and it follows from the proof of Lemma 9.1 in Zhou, van de Geer, and Buhlmann (2009) that $P(\mathcal{T}) \geq 1-(p+q)^{-2}$. Let $\widetilde{\mathbf{Z}} = (\mathbf{X}, \mathbf{e}^{-k})$, $\mathbf{M}^{-k}$ be the submatrix of $\mathbf{M}$ without the $(p+k)$th row and column, $\boldsymbol{\Delta} = n^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} - \mathbf{M}^{-k}$, and

$$\mathcal{Y}_k = \Big\{\max_{j,s}|\boldsymbol{\Delta}_{js}| \leq 8n^{-1/2}(\log(p+q))^{1/2}\Big\}.$$

It then follows from Lemma 9.3 in Zhou, van de Geer, and Buhlmann (2009) that $P(\mathcal{Y}_k) \geq 1-(p+q)^{-2}$. Since $\Lambda_{\min}(d)$ is asymptotically larger than $n^{-1/2}(\log(p+q))^{1/2}$, there exists a constant $c_1 > 0$ such that on the set $\mathcal{T}\cap\mathcal{Y}_k$,

$$\Lambda_{\min}\big(n^{-1}\widetilde{\mathbf{Z}}_{\mathcal{A}_k}^T\widetilde{\mathbf{Z}}_{\mathcal{A}_k}\big) \geq 2c_1\Lambda_{\min}(d),$$

for any subset $\mathcal{A} \subset \{1,\ldots,p+q\}\backslash\{p+k\}$ with $|\mathcal{A}| \leq d$. Furthermore, if $\mathcal{A}_\dagger = \{1 \leq j \leq p : j \in \mathcal{A}\}$ and $\mathcal{A}_\ddagger = \{1 \leq s \leq q : p+s \in \mathcal{A}\}$, then

$$\begin{aligned}
&\big|\Lambda_{\min}(n^{-1}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}) - \Lambda_{\min}(n^{-1}\widetilde{\mathbf{Z}}_{\mathcal{A}}^T\widetilde{\mathbf{Z}}_{\mathcal{A}})\big| \\
&\leq \|n^{-1}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}} - n^{-1}\widetilde{\mathbf{Z}}_{\mathcal{A}}^T\widetilde{\mathbf{Z}}_{\mathcal{A}}\|_2 \leq \|n^{-1}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}} - n^{-1}\widetilde{\mathbf{Z}}_{\mathcal{A}}^T\widetilde{\mathbf{Z}}_{\mathcal{A}}\|_\infty \\
&\leq \|n^{-1}\widehat{\mathbf{y}}_{\mathcal{A}_\ddagger}^T\mathbf{X}_{\mathcal{A}_\dagger}\|_\infty + \|n^{-1}\mathbf{X}_{\mathcal{A}_\dagger}^T\widehat{\mathbf{y}}_{\mathcal{A}_\ddagger}\|_\infty + \|n^{-1}\widehat{\mathbf{y}}_{\mathcal{A}_\ddagger}^T\widehat{\mathbf{y}}_{\mathcal{A}_\ddagger} - n^{-1}\mathbf{e}_{\mathcal{A}_\ddagger}^T\mathbf{e}_{\mathcal{A}_\ddagger}\|_\infty,
\end{aligned}$$

where $\|M\|_2$ is the operator norm of a matrix M, and $\|M\|_\infty = \max_i \sum_j |M_{ij}|$. But since $\widehat{\mathbf{y}}^{-k} = \mathbf{y}^{-k} - \mathbf{X}\,\widehat{\mathbf{B}}^{(0)}_{-k}$ with $\widehat{\mathbf{B}}^{(0)}_{-k}$ the Lasso estimate, we have on the set $\mathcal{T}$,

$$\max(\|n^{-1}\widehat{\mathbf{y}}^T_{\mathcal{A}^\ddagger}\mathbf{X}_{\mathcal{A}^\dagger}\|_\infty, \|n^{-1}\mathbf{X}^T_{\mathcal{A}^\dagger}\widehat{\mathbf{y}}_{\mathcal{A}^\ddagger}\|_\infty) \le O(n^{-1}\lambda_{init}d).$$

Conditional on the set $\mathcal{T}$, it follows from (A1) that

$$\|n^{-1}\widehat{\mathbf{y}}^T_{\mathcal{A}^\ddagger}\widehat{\mathbf{y}}_{\mathcal{A}^\ddagger} - n^{-1}\mathbf{e}^T_{\mathcal{A}^\ddagger}\mathbf{e}_{\mathcal{A}^\ddagger}\|_\infty \le O(n^{-1}d\lambda_{init}n^{-1/2}(\log(p+q))^{1/2}) + O(n^{-2}d\lambda^2_{init}).$$

Therefore, on the set $\mathcal{T}\cap\mathcal{Y}_k$,

$$\begin{aligned}\Lambda_{\min}\big(n^{-1}\mathbf{Z}^T_{\mathcal{A}_k}\mathbf{Z}_{\mathcal{A}_k}\big) &\ge \Lambda_{\min}\big(n^{-1}\widetilde{\mathbf{Z}}^T_{\mathcal{A}_k}\widetilde{\mathbf{Z}}_{\mathcal{A}_k}\big) - \big|\Lambda_{\min}(n^{-1}\mathbf{Z}^T_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}) - \Lambda_{\min}(n^{-1}\widetilde{\mathbf{Z}}^T_{\mathcal{A}}\widetilde{\mathbf{Z}}_{\mathcal{A}})\big| \\ &\ge c_1\Lambda_{\min}(d),\end{aligned}$$

for sufficiently large n, since the cardinality of $\mathcal{A}_k$ is bounded by d and $\Lambda_{\min}(d)$ is asymptotically larger than $\max(n^{-1}\lambda_{init}d, n^{-3/2}\lambda_{init}d(\log(p+q))^{1/2}, n^{-2}\lambda^2_{init}d)$.

It follows from Bickel, Ritov, and Tsybakov (2009) that under Assumption (A2) with $m = d$ and $k_0 = 3$,

$$\begin{aligned}\delta_{\mathcal{A}_k} &:= \max_{j\in\mathcal{A}_k} |\tilde{\zeta}_j - \zeta^*_j| \le 4K(d,d,3,\mathbf{M})^2 n^{-1}d^{1/2}\lambda_{init}; \\ \delta_{\mathcal{A}^c_k} &:= \max_{j\in\mathcal{A}^c_k} |\tilde{\zeta}_j - \zeta^*_j| \le 16K(d,d,3,\mathbf{M})^2 n^{-1}d^{1/2}\lambda_{init},\end{aligned}$$

on set $\mathcal{T}$, where $\tilde{\zeta}_j$ is the solution of the Lasso regression. Therefore,

$$\frac{r_{\min}(\mathcal{A}^c_k)}{r_{\max}(\mathcal{A}_k)} = \frac{\min_{j\in\mathcal{A}_k}|\tilde{\zeta}_j|}{\max_{j\in\mathcal{A}^c_k}|\tilde{\zeta}_j|} \ge \frac{\zeta^*_{\min} - \delta_{\mathcal{A}_k}}{\delta_{\mathcal{A}^c_k}},$$

where $\zeta^*_{\min} = \min_{j\in\mathcal{A}_k}|\zeta^*_j|$. Then it follows from Lemma 10.3 of Zhou, van de Geer, and Buhlmann (2009) that on set $\mathcal{Y}_k$, there exists a positive constant c_2 such that $\|\mathbf{Z}^T_{\mathcal{A}^c_k}\mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}^T_{\mathcal{A}_k}\mathbf{Z}_{\mathcal{A}_k})^{-1}\|_\infty \le c_2 d^{1/2}(\Lambda_{\min}(d))^{-1/2}$. Therefore, on the set $\mathcal{T}\cap\mathcal{Y}_k$, when n is sufficiently large,

$$\|\mathbf{Z}^T_{\mathcal{A}^c_k}\mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}^T_{\mathcal{A}_k}\mathbf{Z}_{\mathcal{A}_k})^{-1}\|_\infty \le \frac{r_{\min}(\mathcal{A}^c_k)}{r_{\max}(\mathcal{A}_k)}(1-\eta),$$

for some $0 < \eta < 1$, provided that $(\zeta^*_{\min})^{-1}n^{-1}d\lambda_{init}(\Lambda_{\min}(d))^{-1/2}K(d,d,3,\mathbf{M})^2 \to 0$.

Finally, from Lemma A.1, for each $k = 1,\ldots,q$,

$$P(\operatorname{sgn}(\widehat{\zeta}) \ne \operatorname{sgn}(\zeta^*)) = O((p+q)^{-2}),$$

provided $n^{-1}d\lambda_{init} \to 0$ and $\min\big(n^{-1}\lambda^2_n r^2_{\min}(\mathcal{A}^c_k), n\Lambda_{\min}(d)(\zeta^*_{\min})^2\big)(\log(p+q))^{-1} \to \infty$. Consequently, $P(\operatorname{sgn}(\widehat{\mathbf{B}}) \ne \operatorname{sgn}(\mathbf{B}) \text{ or } \operatorname{sgn}(\widehat{\mathbf{\Omega}}) \ne \operatorname{sgn}(\mathbf{\Omega})) \le qO((p+q)^{-2})$, which implies the desired result.

Lemma A.1. *Consider* (A.3), *where the design matrix* $\mathbf{Z}$ *satisfies*

$$\Lambda_{\min}\big(n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\big) \geq c_1\Lambda_{\min}(d) > 0, \tag{A.4}$$

$$\|\mathbf{Z}_{\mathcal{A}_k^c}^T\mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\|_\infty \leq \frac{r_{\min}(\mathcal{A}_k^c)}{r_{\max}(\mathcal{A}_k)}(1-\eta) \tag{A.5}$$

for some constants $c_1 > 0$ *and* $0 < \eta < 1$, $r_{\min}(\mathcal{A}_k^c) = \min_{j\in\mathcal{A}_k^c} r_j$, *and* $r_{\max}(\mathcal{A}_k) = \max_{j\in\mathcal{A}_k} r_j$. *If* $\zeta^*_{\min} = \min_j |\zeta_j^*|$ *is asymptotically larger than* $(\Lambda_{\min}(d))^{-1} O\big(\max\big(n^{-1}\lambda_n d^{1/2} r_{\max}(\mathcal{A}_k),\ n^{-1}d^{1/2}\lambda_{init},\ n^{-1/2}d^{1/2}(\log(p+q))^{1/2},\ n^{-2}d\lambda_{init}^{-2} K(d,d,3,\mathbf{M})\big)\big)$, *then* $P(\operatorname{sgn}(\widehat{\zeta}) \neq \operatorname{sgn}(\zeta^*)) = O((p+q)^{-2})$, *provided* $n^{-1/2}d\lambda_{init} \to 0$ *and* $\min\big(n^{-1}\lambda_n^2 r_{\min}^2(\mathcal{A}_k^c), n\Lambda_{\min}(d)(\zeta^*_{\min})^2\big)(\log(p+q))^{-1} \to \infty$.

Proof of Lemma A.1. Let $\mathbf{z}^j$ be the jth column of $\mathbf{Z}$. It follows from the Karush-Kuhn-Tucker condition that $\widehat{\zeta}$ must satisfy

$$(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}\zeta) = \lambda_n r_j \operatorname{sign}(\zeta_j),\ \text{if } \widehat{\zeta}_j \neq 0; \tag{A.6}$$

$$\big|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}\zeta)\big| \leq \lambda_n r_j,\ \text{if } \zeta_j = 0. \tag{A.7}$$

Consider the equation based on $\mathbf{Z}_{\mathcal{A}_k}$,

$$\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\bar{\zeta}_{\mathcal{A}_k} = \lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k},$$

where $\bar{\mathbf{s}}_{\mathcal{A}_k} = \big(r_j \operatorname{sign}(\zeta_j^*); j \in \mathcal{A}_k\big)$. By (A.2), it has the solution

$$\bar{\zeta}_{\mathcal{A}_k} = \zeta^*_{\mathcal{A}_k} + (\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}(\mathbf{Z}_{\mathcal{A}_k}^T(\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_k) - \lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}). \tag{A.8}$$

If $\operatorname{sgn}(\bar{\zeta}_{\mathcal{A}_k}) = \operatorname{sgn}(\zeta^*_{\mathcal{A}_k})$, $\widehat{\zeta}$ with $(\widehat{\zeta}_j)_{j\in\mathcal{A}_k} = \bar{\zeta}_{\mathcal{A}_k}$, $(\widehat{\zeta}_j)_{j\notin\mathcal{A}_k} = 0$ is a solution of (A.6)-(A.7). Therefore, $\operatorname{sgn}(\widehat{\zeta}) = \operatorname{sgn}(\zeta^*)$ if

$$\operatorname{sgn}(\bar{\zeta}_{\mathcal{A}_k}) = \operatorname{sgn}(\zeta^*_{\mathcal{A}_k}),\ \text{and } \big|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\bar{\zeta}_{\mathcal{A}_k})\big| \leq \lambda_n r_j,\ \text{if } j \notin \mathcal{A}_k.$$

This statement is similar to Proposition 1 of Zhao and Yu (2006) and (S.5) of Huang, Ma, and Zhang (2008b). It implies that

$$\begin{aligned} P(\operatorname{sgn}(\widehat{\zeta}) \neq \operatorname{sgn}(\zeta^*)) &\leq P(\operatorname{sgn}(\zeta_{\mathcal{A}_k}) \neq \operatorname{sgn}(\zeta^*_{\mathcal{A}_k})) \\ &\quad + P(|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\bar{\zeta}_{\mathcal{A}_k})| > \lambda_n r_j, \exists\ j \notin \mathcal{A}_k). \end{aligned}$$

We now bound the probabilities on the right hand side conditional on the set $\mathcal{T}$. For brevity, we use $P(\cdot)$ to denote the conditional probability given $\mathcal{T}$ in the remainder of the proof. By (A.8),

$$\begin{aligned} P(\operatorname{sgn}(\bar{\zeta}_{\mathcal{A}_k}) \neq \operatorname{sgn}(\zeta^*_{\mathcal{A}_k})) &\leq P\Big(|\zeta_j^* - \bar{\zeta}_j| \geq |\zeta_j^*|, \exists\ j \in \mathcal{A}_k\Big), \\ &\leq P\Big(\big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\epsilon}_k\big| \geq \zeta^*_{\min}/2\Big) \\ &\quad + P\Big(\big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k\big| + \big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}\big| \geq \zeta^*_{\min}/2\Big), \end{aligned}$$

where $\mathbf{1}_j$ is a vector of zeros except the jth component being 1 and, by (A.4),

$$\begin{aligned}\big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}\big| &\le \big(c_1\Lambda_{\min}(d)\big)^{-1}\|n^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}\|\\ &\le \big(c_1\Lambda_{\min}(d)\big)^{-1}n^{-1}\lambda_n d^{1/2}r_{\max}(\mathcal{A}_k).\end{aligned}$$

From the definition of $\mathcal{T}$ and the initial Lasso estimates, that

$$\begin{aligned}&\big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k\big| \le \|(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k\|\\ &\quad\le \big(c_1\Lambda_{\min}(d)\big)^{-1}\|n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k\| \le \big(c_1\Lambda_{\min}(d)\big)^{-1}\big(\|n^{-1}\mathbf{X}_{\mathcal{A}_k^\beta}^T\boldsymbol{\xi}_k\| + \|n^{-1}\widehat{\mathbf{y}}_{\mathcal{A}_k^\omega}^T\boldsymbol{\xi}_k\|\big)\\ &\quad\le (c_1\Lambda_{\min}(d))^{-1}\big(O(n^{-1}d^{1/2}\lambda_{init}) + O(n^{-1/2}d^{1/2}(\log(p+q))^{1/2})\\ &\qquad +O(n^{-2}d\lambda_{init}^2K(d,d,3,\mathbf{M}))\big).\end{aligned}$$

Since $\zeta_{\min}^*/2$ is asymptotically larger than the upper bounds in the last two inequalities and

$$\begin{aligned}n^{-1}\|\mathbf{1}_j^T(n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\| &\le n^{-1/2}(\Lambda_{\min}(n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}))^{-1/2}\\ &\le n^{-1/2}\big(c_1\Lambda_{\min}(d)\big)^{-1/2},\end{aligned}$$

there exists some positive constant c_3 such that for sufficiently large n,

$$\begin{aligned}P(\operatorname{sgn}(\bar{\zeta}_{\mathcal{A}_k}) \ne \operatorname{sgn}(\zeta_{\mathcal{A}_k}^*)) &\le P\Big(\big|\mathbf{1}_j^T(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\epsilon}_k\big| \ge \frac{1}{2}\zeta_{\min}^*, \exists\, j\in\mathcal{A}_k\Big)\\ &\le c_3 d\exp\Big(-nc_1\Lambda_{\min}(d)\frac{(\zeta_{\min}^*)^2}{4\tilde{\sigma}_{kk}^*}\Big).\end{aligned}$$

To bound $P(|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\bar{\zeta})| > \lambda_n r_j, \exists\, j\notin\mathcal{A}_k)$, we have

$$\begin{aligned}&P(|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\bar{\zeta})| > \lambda_n r_j, \exists\, j\notin\mathcal{A}_k)\\ &\quad= P\Big(|(\mathbf{z}^j)^T(\mathbf{H}_{\mathcal{A}_k}(\boldsymbol{\xi}_k+\boldsymbol{\epsilon}_k) + \mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k})| > \lambda_n r_j, \exists\, j\notin\mathcal{A}_k\Big)\\ &\quad\le P\Big(|(\mathbf{z}^j)^T\mathbf{H}_{\mathcal{A}_k}(\boldsymbol{\xi}_k+\boldsymbol{\epsilon}_k)| + |(\mathbf{z}^j)^T\mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}| \ge \lambda_n r_j, \exists\, j\notin\mathcal{A}_k\Big),\end{aligned}$$

where $\mathbf{H}_{\mathcal{A}_k} = \mathbf{I} - \mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\mathbf{Z}_{\mathcal{A}_k}^T$. Now (A.5) implies that $|(\mathbf{z}^j)^T\mathbf{Z}_{\mathcal{A}_k}(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k})^{-1}\lambda_n\bar{\mathbf{s}}_{\mathcal{A}_k}| \le \lambda_n r_j(1-\eta)$ for any $j\in\mathcal{A}_k^c$, and therefore,

$$\begin{aligned}&P(|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\zeta)| > \lambda_n r_j, \exists\, j\notin\mathcal{A}_k)\\ &\quad\le P\Big(|(\mathbf{z}^j)^T\mathbf{H}_{\mathcal{A}_k}(\boldsymbol{\xi}_k+\boldsymbol{\epsilon}_k)| \ge \eta\lambda_n r_j, \exists\, j\notin\mathcal{A}_k\Big).\end{aligned}$$

By Lemma 11.3 of Zhou, van de Geer, and Buhlmann (2009) and the fact that $\sigma_{jj}^* = 1$, there exists a positive constant c_4 such that $P(\max_j n^{-1}(\mathbf{z}^j)^T\mathbf{z}^j \ge c_4) \le (p+q)^{-2}$. Conditional on the set $\{\max_j n^{-1}(\mathbf{z}^j)^T\mathbf{z}^j \le c_4\}$, $\|(\mathbf{z}^j)^T\mathbf{H}_{\mathcal{A}_k}\| \le$

$(c_4 n)^{1/2}$, $\|\boldsymbol{\xi}_k\| \le O(n^{-1/2} d\lambda_{init})$ by (A1). Since $n^{-1/2} d\lambda_{init} = o(1)$, there exists some positive constant c_5 such that

$$P(|(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}_{\mathcal{A}_k}\zeta)| > \lambda_n r_j, \exists\, j \notin \mathcal{A}_k) \le c_5(p+q)\exp\left(-\frac{\eta^2\lambda_n^2 r_{\min}^2(\mathcal{A}_k^c)}{2c_4 n\tilde{\sigma}_{kk}^*}\right).$$

Combining these results, for sufficiently large n,

$$\begin{aligned} P(\operatorname{sgn}(\widehat{\zeta}) \ne \operatorname{sgn}(\zeta^*)) &\\ &\le (p+q)^{-2} + c_3 d\exp\left(-\frac{nc_1\Lambda_{\min}(d)(\zeta_{\min}^*)^2}{2\tilde{\sigma}_{kk}^*}\right)\\ &\quad + c_5(p+q)\exp\left(-\frac{\eta^2\lambda_n^2 r_{\min}^2(\mathcal{A}_k^c)}{2c_4 n\tilde{\sigma}_{kk}^*}\right), \end{aligned}$$

and the desired result follows.

Proof of Theorem 2. The solution of (3.3) is the same as that of (A.3), where $\widehat{\zeta} = (\widehat{\beta}_k^T, \widehat{\gamma}_k^T)^T$ and satisfies that $-2(\mathbf{z}^j)^T(\mathbf{y}^k - \mathbf{Z}\widehat{\zeta}) + \lambda_n r_j \operatorname{sign}(\widehat{\zeta}_j) = 0$, for any $j \in \mathcal{A}_k$. Let $\widehat{\mathbf{s}}_{\mathcal{A}_k} = (r_j \operatorname{sign}(\widehat{\zeta}_j);\ j \in \mathcal{A}_k)$, then $-2\mathbf{Z}_{\mathcal{A}_k}^T(\mathbf{y}^k - \mathbf{Z}\widehat{\zeta}) + \lambda_n \widehat{\mathbf{s}}_{\mathcal{A}_k} = 0$, or equivalently,

$$\frac{1}{\sqrt{n}}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}(\widehat{\zeta}_{\mathcal{A}_k} - \zeta_{\mathcal{A}_k}^*) = \frac{1}{\sqrt{n}}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\epsilon}_k + \frac{1}{\sqrt{n}}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k - \frac{\lambda_n}{2\sqrt{n}}\widehat{\mathbf{s}}_{\mathcal{A}_k} - \frac{1}{\sqrt{n}}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k^c}\widehat{\zeta}_{\mathcal{A}_k^c}.$$

By the proof of Theorem 1, on the set $\mathcal{T} \cap \mathcal{Y}_k$, $\Lambda_{\min}(n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}) \ge c_1\Lambda_{\min}(d) > 0$. If $\boldsymbol{\Sigma}_k = n^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}$, on the set $\mathcal{T} \cap \mathcal{Y}_k$, for any $|\mathcal{A}_k| \times 1$ vector α,

$$\begin{aligned} \sqrt{n}s_k^{-1}\alpha^T(\widehat{\zeta}_{\mathcal{A}_k} - \zeta_{\mathcal{A}_k}^*) &= \frac{1}{\sqrt{n}}s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\epsilon}_k + \frac{1}{\sqrt{n}}s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\boldsymbol{\xi}_k\\ &\quad - \frac{\lambda_n}{2\sqrt{n}}s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\widehat{\mathbf{s}}_{\mathcal{A}_k} - \frac{1}{\sqrt{n}}s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k^c}\widehat{\zeta}_{\mathcal{A}_k^c}. \end{aligned}$$

We show that on the set $\mathcal{T} \cap \mathcal{Y}_k$ the last three components converge to 0 in probability uniformly with respect to α. First, the proof of Theorem 1 implies that $P(\widehat{\zeta}_{\mathcal{A}_k^c} = 0) \ge 1 - (p+q)^{-2} \to 1$, and thus

$$P\left(\frac{1}{\sqrt{n}}s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k^c}\widehat{\zeta}_{\mathcal{A}_k^c} = 0\right) \longrightarrow 1.$$

Second, by (A3) and the fact that $\|\alpha\| = 1$,

$$\begin{aligned} \left|\frac{1}{\sqrt{n}}\lambda_n s_k^{-1}\alpha^T\boldsymbol{\Sigma}_k^{-1}\widehat{\mathbf{s}}_{\mathcal{A}_k}\right| &\le \frac{1}{\sqrt{n}}\lambda_n s_k^{-1}(\Lambda_{\min}(\boldsymbol{\Sigma}_k))^{-1}\|\alpha\|\|\widehat{\mathbf{s}}_{\mathcal{A}_k}\|\\ &\le \frac{1}{\sqrt{n}}\lambda_n s_k^{-1}(c_1\Lambda_{\min}(d))^{-1}d^{1/2}(\zeta_{\min}^*)^{-1} \longrightarrow 0, \end{aligned}$$

except on an event with probability tending to zero. Third, on the set $\mathcal{T} \cap \mathcal{Y}_k$,

$$\begin{aligned}\Big|\frac{1}{\sqrt{n}} s_k^{-1} \alpha^T \mathbf{\Sigma}_k^{-1} \mathbf{Z}_{\mathcal{A}_k}^T \boldsymbol{\xi}_k\Big| &\le s_k^{-1} \left\| \frac{1}{\sqrt{n}} \alpha^T \mathbf{\Sigma}_k^{-1} \mathbf{Z}_{\mathcal{A}_k}^T \right\| \|\boldsymbol{\xi}_k\| = s_k^{-1} (\alpha^T \mathbf{\Sigma}_k^{-1} \alpha)^{1/2} \|\boldsymbol{\xi}_k\| \\ &\le s_k^{-1} (c_1 \Lambda_{\min}(d))^{-1/2} \|\boldsymbol{\xi}_k\| \longrightarrow 0,\end{aligned}$$

where $\|\boldsymbol{\xi}_k\| \le O(n^{-1/2} d \lambda_{init})$, as in the proof of Theorem 1.

Therefore, on the set $\mathcal{T} \cap \mathcal{Y}_k$,

$$\sqrt{n} s_k^{-1} \alpha^T (\widehat{\zeta}_{\mathcal{A}_k} - \zeta^*_{\mathcal{A}_k}) = \frac{1}{\sqrt{n}} s_k^{-1} \alpha^T \mathbf{\Sigma}_k^{-1} \mathbf{Z}_{\mathcal{A}_k}^T \boldsymbol{\epsilon}_k + o_p(1),$$

where $(1/\sqrt{n}) s_k^{-1} \alpha^T \mathbf{\Sigma}_k^{-1} \mathbf{Z}_{\mathcal{A}_k}^T \boldsymbol{\epsilon}_k \xrightarrow{d} N(0,1)$ by verifying the conditions of the Lindeberg-Feller Central Limit Theorem as in Huang, Horowitz, and Ma (2008a). Furthermore, on the set $(\mathcal{T} \cap \mathcal{Y}_k)^c$, $|\sqrt{n} s_k^{-1} \alpha^T (\widehat{\zeta}_{\mathcal{A}_k} - \zeta^*_{\mathcal{A}_k})| \le s_k^{-1} O_p((p+q+d\lambda_n))$ by Theorem 1 of Huang, Horowitz, and Ma (2008a), and $P((\mathcal{T} \cap \mathcal{Y}_k)^c) \le (p+q)^{-2}$, by the proof of Theorem 1. As $d\lambda_n = o(p+q)$, the desired asymptotic normality follows.

References

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression. *J. Roy. Statist. Soc. Ser. B* **59**, 3-54.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **208**, 759-771.

Chen, L. and Huang, J. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.* **107**, 1533-1545.

Edwards, D. (2000). *Introduction to Graphical Modeling*. 2nd edition. Springer, New York.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Statist.* **3**, 521-541.

Finegold, M. and Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Statist.* **5**, 1057-1080.

Friedman, J., Hastie,T. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302-332.

Friedman, J., Hastie,T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

Huang, J., Horowitz, J. and Ma, S. (2008a). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Huang, J., Ma, S. and Zhang, C.-H. (2008b). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

Kendziorski, C., Chen, M., Yuan, M., Lan, H. and Attie, A. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19-27.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate Anal.* **111**, 241-255.

Lee, W., Du, Y., Sun, W., Heyes, D. and Liu, Y. (2012). Multiple response regression for Gaussian mixture models with known labels. *Statist. Anal. and Data Mining* **5**, 493-508.

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7**, 302-317.

Meinshausen, N. (2007). Lasso with relaxation. *Comput. Statist. Data Anal.* **52**, 374-393.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104**, 735-746.

Rothman, A., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19**, 947-962.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

TCGA (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068.

Turlach, B., Venables, W. and Wright, S. (2005). Simultaneous variable selection. *Technometrics* **47**, 349-363.

Verhaak, R., Hoadley, K., Purdom, E., Wang, V., Qi, Y., Wilkerson, M., Miller, C., Ding, L., Golub, T., Mesirov, J., Alexe, G., Lawrence, M., OKelly, M., Tamayo, P., Weir, B., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H., Hodgson, J., James, C., Sarkaria, J., Brennan, C., Kahn, A., Spellman, P., Wilson, R., Speed, T., Gray, J., Meyerson, M., Getz, G., Perou, C. and Hayes, D. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, **17**, 98-110.

Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Statist.* **5**, 2630-2650.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Statist. Soc. Ser. B* **69**, 329-346.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Zhou, S., van de Geer, S. and Buhlmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. Manuscript, arxiv:0903.2515.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago.

Department of Mathematics, City University of Hong Kong.

E-mail: junhui@uic.edu

(Received July 2013; accepted May 2014)