Statistica Sinica **24** (2014), 1633-1654
doi:http://dx.doi.org/10.5705/ss.2013.019

VARIABLE SELECTION FOR HIGH DIMENSIONAL MULTIVARIATE OUTCOMES

Tamar Sofer, Lee Dicker and Xihong Lin

University of Washington, Rutgers University and Harvard school of Public Health

Abstract: We consider variable selection for high-dimensional multivariate regression using penalized likelihoods when the number of outcomes and the number of covariates might be large. To account for within-subject correlation, we consider variable selection when a working precision matrix is used and when the precision matrix is jointly estimated using a two-stage procedure. We show that under suitable regularity conditions, penalized regression coefficient estimators are consistent for model selection for an arbitrary working precision matrix, and have the oracle properties and are efficient when the true precision matrix is used or when it is consistently estimated using sparse regression. We develop an efficient computation procedure for estimating regression coefficients using the coordinate descent algorithm in conjunction with sparse precision matrix estimation using the graphical LASSO (GLASSO) algorithm. We develop the Bayesian Information Criterion (BIC) for estimating the tuning parameter and show that BIC is consistent for model selection. We evaluate finite sample performance for the proposed method using simulation studies and illustrate its application using the type II diabetes gene expression pathway data.

Key words and phrases: BIC, consistency, correlation, efficiency, model selection, multiple outcomes, oracle estimator.

1. Introduction

Correlated multivariate responses are often observed in health science studies where the number of responses per subject may be large and correlated, and numerous covariates may be observed for each subject. For instance, in genetic pathway studies, one is often interested in associating gene expressions in a genetic pathway to such clinical covariates as exposure, treatment, and individual characteristics. A pathway may have tens and hundreds of genes and only a small number of gene expressions are likely to be associated with exposures. A question of particular interest is to identify a subset of genes that are associated with exposures while accounting for the fact that the gene expressions with the same pathway are likely to be correlated and the number of genes is large relative to the sample size. In this paper, we consider the problem of variables selection in the presence of multivariate outcomes. We allow the numbers of outcomes and covariates to be large and the outcomes to be correlated.

There is a vast literature on variable selection for independent data with methods often based on penalized likelihoods. Examples include LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), adaptive LASSO (Zou (2006)), SELO (Dicker, Huang, and Lin (2013)) and MC+ (Zhang (2010)). A desirable property for variable selection is the oracle property (Fan and Li (2001)). It is known that LASSO is often not consistent for variable selection, while SCAD, Adaptive LASSO, SELO and MC+ all enjoy the oracle property. These methods have been extended to the situations where the number of independent variables increases with sample size (Fan and Peng (2004)). Tuning parameter selection is important for variable selection. BIC has been proposed for selecting the tuning parameter and has been shown to be consistent for variable selection (Wang, Li, and Leng (2009); Dicker, Huang, and Lin (2013)).

Compared to independent data, a challenge in variable selection for multivariate outcomes is that multiple outcomes of each subject are generally correlated. If outcomes have the same unknown mean, sparse precision matrix (inverse covariance matrix) estimation has been a fertile research area. For instance, Schäfer and Strimmer (2005) proposed shrinkage estimators, while Meinshausen and Buhlmann (2006) and Yuan and Lin (2007) proposed Lasso penalized estimators for the covariance or precision matrix. Lam and Fan (2009) established asymptotic theory for inverse covariance estimation when the outcomes have mean zero. In general, this work dealt with mean zero outcome variables, and did not consider the effect of diverging number of mean parameters on the estimation of the precision matrix.

Limited research has been done on variable selection for regression coefficients for multiple outcomes. Several authors have proposed methods for variable selection under the assumption that a covariate either affects all or none of the outcomes predictors (Brown, Frean, and Vannucci (1999); Turlach, Venables, and Wright (2005); Peng et al. (2010)). These authors implicitly assumed working independence among outcomes by ignoring between-outcome correlation in their procedure, and did not present large sample properties of the resulting estimators. Rothman et al. (2010) proposed an iterative algorithm for variable selection and estimation in high-dimensional multivariate regression that alternately estimates the inverse covariance matrix and regression parameters under the ℓ_1 penalty. Asymptotic properties of these estimators were not provided.

We develop variable selection and estimation procedures for regression coefficients in a multivariate regression model when the number of outcomes and the number of regression coefficients are likely to be large. We first propose variable selection for a given working precision matrix, that allows the within-subject correlation to be misspecified. We then propose a "two-stage" estimation procedure

to jointly estimate regression coefficients and the precision matrix using penalized likelihood methods, that results in efficiency gains in regression coefficient estimators.

We focus on variable selection for regression coefficients in multivariate regression while most existing literature has focused on covariance estimation. Second, unlike Rothman et al. (2010) who primarily focused on the LASSO penalty for variable selection of regression coefficients, we can accommodate a wide range of concave penalty functions. We show that our regression parameter estimator is consistent for an arbitrary working precision matrix, has the oracle property, if an appropriate penalty is used, when the true precision matrix is used or is consistently estimated, and that the number of parameters might diverge to ∞. We propose a BIC criterion for tuning parameter selection in the multivariate penalized regression problem, and show that it is consistent for selecting the true regression model even if the number of parameters diverges. Our "two-stage" algorithm is computationally more efficient than the MRCE algorithm introduced in Rothman et al. (2010).

The rest of the paper is organized as follows. We describe the model in Section 2. In Section 3.1, we discuss sparse multivariate regression estimation using an arbitrary working precision matrix. In Section 3.1.1, we propose a two-stage joint estimation procedure for regression parameters and the precision matrix, and a computationally efficient estimation algorithm. We study its properties and show model selection consistency can be achieved using BIC. In Section 6, we present simulation studies that evaluate the finite sample performance of the proposed methods, compare the use of different penalties in penalized multivariate regression, and demonstrate the performance of the BIC. In Section 7, we apply the proposed method to the analysis of type II diabetes gene expression data, and there is a discussion in Section 8.

2. The Model

Consider a multivariate regression model for the ith subject

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ are m-dimensional vectors of outcomes and (unobserved) errors, respectively, $\mathbf{X}_i = \boldsymbol{x}_i^T \otimes \mathbf{I}_m$ is a $m \times p$ dimensional matrix of covariates, where $p = mp_0$, and $\boldsymbol{x}_i$ is a $p_0 \times 1$ covariate vector. Let $\boldsymbol{\beta}_{ok}$ be a vector of p_0-dimensional regression parameters for the k^{th} outcome, so $(\boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{0m}^T)^T = \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional vector of unknown regression parameters. We assume that $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ are iid, with $E(\boldsymbol{\epsilon}_i) = 0$ and $\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_m$, for some unknown $m \times m$ covariance matrix $\boldsymbol{\Sigma}_m$. The total number of subjects is n and

the total number of observations is $N = nm$. We will take $n \to \infty$. Our asymptotic framework will also allow $m \to \infty$ and $p_0 \to \infty$. We first require in Section 4 that p grows more slowly than n, and then extend the asymptotic results to $p > n$ in Section 5. Since p changes with n, it is implicit that $\boldsymbol{\beta}$ may vary with n as well.

Let $A = \{j;\ \beta_j \neq 0\}$ be the subset of $\{1, \ldots, p\}$ corresponding to the nonzero entries of $\boldsymbol{\beta}$. We sometimes refer to A as the "true model." Let $\mathbf{Y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T \in \mathbb{R}^N$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_n^T)^T \in \mathbb{R}^N$, then (2.1) may be rewritten as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $N \times 1$ vectors, $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\mathbf{X}$ is a $N \times p$ matrix. Let $\boldsymbol{\Sigma}$ be the $N \times N$ block diagonal matrix whose i-th diagonal component is $\boldsymbol{\Sigma}_m$ $(1 \leq i \leq n)$. We write the true precision matrix $\boldsymbol{\Omega}_m = (\omega_{ij}) = \boldsymbol{\Sigma}_m^{-1}$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

If $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_m)$, the joint log-likelihood for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}_m$ is, apart from a constant,

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\Omega}_m) &= \ln|\boldsymbol{\Omega}_m| - \frac{1}{n}\sum_{i=1}^{N}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Omega}_m(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= \ln|\boldsymbol{\Omega}_m| - \operatorname{tr}\left\{\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\beta})\boldsymbol{\Omega}_m\right\}, \end{aligned} \tag{2.2}$$

where $\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T$. We do not need to assume that the $\boldsymbol{\epsilon}_i$ are normally distributed, but (2.2) is basic to our method.

3. Estimation Procedures for Variable Selection

In Section 3.1, we propose an estimation procedure for variable selection given a working precision matrix that might be misspecified; in Section 3.1.1 we introduce joint estimation of the regression parameters and the precision matrix; in Section 3.1.2, we propose a two-stage joint estimation procedure.

3.1. Estimation of regression parameters using a working precision matrix

Let

$$Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m) = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Lambda}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\sum_{j=1}^{p}P_\lambda(|\beta_j|), \tag{3.1}$$

where $P_\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ is a concave and differentiable penalty function indexed by a nonnegative tuning parameter $\lambda \geq 0$ satisfying $P_\lambda^{(1)}(0) = 0$, with $\boldsymbol{\Lambda}$ a working precision matrix that might misspecify the true precision matrix $\boldsymbol{\Omega}$. This objective function might be referred to as a penalized general likelihood. By transforming outcomes and covariates using $\tilde{\mathbf{Y}} = \boldsymbol{\Lambda}^{1/2}\mathbf{Y}$ and $\tilde{\mathbf{X}} = \boldsymbol{\Lambda}^{1/2}\mathbf{X}$, we can minimize $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m) = n^{-1}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + 2\sum_{j=1}^{p}P_\lambda(|\beta_j|)$ by the penalized likelihood methods for independent data.

Several penalty functions have been proposed for variable selection and estimation for independent data. Examples are as follows:

(P1) The L_1 penalty, $P_\lambda(|\theta|) = \lambda|\theta|$, also known as the Lasso (Tibshirani (1996)).

(P2) The Adaptive Lasso, $P_\lambda(|\theta|) = \lambda w^{-1}|\theta|$ for some data-dependent weight w (Zou (2006)).

(P3) The SCAD penalty has $P_\lambda(0) = 0$ and the derivative

$$P'_\lambda(|\theta|) = \lambda\Big\{I(|\theta| \le \lambda) + \frac{(a\lambda - |\theta|)^+}{(a-1)\lambda} I(|\theta| > \lambda)\Big\},$$

where $a > 0$ is another tuning parameter often taken to be $a = 3.7$ (Fan and Li (2001)).

(P4) The seamless-L_0 (SELO) penalty

$$P_\lambda(\theta) = \frac{\lambda}{\log(2)} \log\Big\{\frac{|\theta|}{|\theta| + \tau} + 1\Big\},$$

where $\tau > 0$ is another tuning parameter often taken to be $\tau = 0.01$ when x's are standardized (Dicker, Huang, and Lin (2013)).

All these penalties satisfy $\lim_{|\theta|\to 0} P'_\lambda(|\theta|) > 0$ provided $\lambda > 0$, a property that ensures sparsity of the estimated parameters.

To estimate standard errors for the regression parameters, we used a quadratic approximation of the penalty function (Fan and Li (2001)). Standard errors are estimated only for the parameters that are estimated as non-zeros, and are set to be zero if the estimates are zero. Let $\hat{A}$ be a set of indices with estimated non-zero parameters. Let $\boldsymbol{\beta}_{\hat{A}}$ be the parameter sub-vector and $\mathbf{X}_{\hat{A}}$ be the covariate sub-matrix that corresponds to these indices. Given λ, one can show that the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}_{\hat{A}}$ is given by the sandwich estimator (Liang and Zeger (1986))

$$\begin{aligned}\widehat{cov}(\widehat{\boldsymbol{\beta}}_{\hat{A}}) = &\Big\{\mathbf{X}_{\hat{A}}^T \boldsymbol{\Lambda} \mathbf{X}_{\hat{A}} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_{\hat{A}})\Big\}^{-1} \Big\{\sum_{i=1}^n \mathbf{X}_{\hat{A},i}^T \boldsymbol{\Lambda} \widehat{\boldsymbol{\Sigma}}_i \boldsymbol{\Lambda} \mathbf{X}_{\hat{A},i}\Big\} \\ &\times \Big\{\mathbf{X}_{\hat{A}}^T \boldsymbol{\Lambda} \mathbf{X}_{\hat{A}} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_{\hat{A}})\Big\}^{-1}, \end{aligned} \tag{3.2}$$

where $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Sigma}}_i(\widehat{\boldsymbol{\beta}}_{\hat{A}}) = (\boldsymbol{y}_i - \mathbf{X}_{\hat{A},i}\widehat{\boldsymbol{\beta}}_{\hat{A}})(\boldsymbol{y}_i - \mathbf{X}_{\hat{A},i}\widehat{\boldsymbol{\beta}}_{\hat{A}})^T$ and, for a set of indices $\mathcal{K}$ of size k, $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_\mathcal{K}) = \mathrm{diag}\{P'_\lambda(|\beta_{\mathcal{K},(1)}|)/|\beta_{\mathcal{K},(1)}|, \ldots, P'_\lambda(|\beta_{\mathcal{K},(k)}|)/|\beta_{\mathcal{K},(k)}|\}$.

3.1.1. Joint estimation of regression coefficients and precision matrix

Results in Section 4.1 show that misspecification of the precision matrix leads to less efficient estimators of regression coefficients. To improve the efficiency, we propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ simultaneously. It is well known that although the MLE of the error covariance matrix

$$\mathbf{S} = \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^T$$

is consistent, it is not positive definite when the number of outcomes m is larger than the number of observations, $m > n$, and even if $m < n$ and the number of parameters is large, the MLE is unstable (Schäfer and Strimmer (2005)). Likewise the estimator of the precision matrix $\mathbf{S}^{-1}$, is unstable, if it exists. This motivates regularized estimation of the precision matrix $\boldsymbol{\Omega}$ as well as of $\boldsymbol{\beta}$. Let $P_\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ be a concave, differentiable penalty function, $\gamma \geq 0$, $P_\gamma(0) = 0$. We jointly estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ by minimizing the objective function

$$Q(\boldsymbol{\beta}, \boldsymbol{\Omega}_m) = -\ln|\boldsymbol{\Omega}_m| + \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\beta})\boldsymbol{\Omega}_m\right] + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|) + 2\sum_{1\leq i<j\leq m} P_\gamma(|\omega_{ij}|),$$

where $P_\gamma^{(2)}(\cdot)$ is a penalty for $\boldsymbol{\Omega}_m$.

3.1.2. Two-stage joint estimation

Minimizing $Q(\boldsymbol{\beta}, \boldsymbol{\Omega})$ simultaneously over $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ is computationally difficult as the number of parameters is very large. As an alternative, consider an objective function for estimation of $\boldsymbol{\Omega}$ given $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\Omega}|\boldsymbol{\beta}) = -\ln|\boldsymbol{\Omega}_m| + \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\beta})\boldsymbol{\Omega}_m\right] + 2\sum_{1\leq i<j\leq m} P_\gamma(|\omega_{ij}|) \tag{3.3}$$

to maximize (3.3) to obtain an estimator of $\boldsymbol{\Omega}$ if the L_1 penalty is used. We propose a two-stage procedure for estimation of $\boldsymbol{\Omega}$ by iteratively minimizing $Q(\boldsymbol{\Omega}|\widehat{\boldsymbol{\beta}})$ in (3.3) by setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, and estimating $\boldsymbol{\beta}$ by minimizing $Q(\boldsymbol{\beta}|\widehat{\boldsymbol{\Omega}})$ in (3.1) by setting $\boldsymbol{\Lambda} = \widehat{\boldsymbol{\Omega}}$.

Suppose the tuning parameters λ and γ are known. At the first stage, using the results in Section 3.1, we estimate a consistent estimator $\hat{\boldsymbol{\beta}}^{(1)}$ by assuming a working independent precision matrix; we then estimate $\boldsymbol{\Omega}$ using the obtained $\widehat{\boldsymbol{\beta}}^{(1)}$ in minimization of $Q(\boldsymbol{\Omega}|\widehat{\boldsymbol{\beta}})$ and call this estimator $\hat{\boldsymbol{\Omega}}^{(1)}$. At a second stage, we improve efficiency of estimation of $\boldsymbol{\beta}$ using the consistently-estimated $\widehat{\boldsymbol{\Omega}}^{(1)}$ from stage 1 and call this estimator $\widehat{\boldsymbol{\beta}}$. We further improve upon estimation of $\boldsymbol{\Omega}$ by using the efficient estimate of $\widehat{\boldsymbol{\beta}}$. Hence, at the first stage consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are calculated; at the second, their efficient counterparts are obtained.

The tuning parameters are not known and can be estimated e.g, using BIC or data validation procedure (see Section 3.2). We propose a two-stage joint estimation procedure that is computationally efficient and show that it has desirable properties.

Stage 1. Set $\widehat{\mathbf{\Omega}}_m^{(0)} = \mathbf{I}_m$.

- Let $\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{(1)}(\hat{\lambda})$ be a model selection consistent working independence estimator of $\boldsymbol{\beta}$, where the penalty satisfies the regularity conditions in Section 3, and $\hat{\lambda}$ is estimated by some selection criterion, see Section 3.2.
- Set $\widehat{\mathbf{\Omega}}^{(1)} = \widehat{\mathbf{\Omega}}(\hat{\gamma})$ to be the sparse consistent estimator of $\mathbf{\Omega}$ given $\widehat{\boldsymbol{\beta}}^{(1)}$ by minimizing $Q(\mathbf{\Omega}|\widehat{\boldsymbol{\beta}}^{(1)})$, where the tuning parameter $\hat{\gamma} = \hat{\gamma}(\hat{\lambda})$ is estimated using a selection criterion, see Section 3.2.

Stage 2. For each tuning parameter λ:

- Set $\widehat{\boldsymbol{\beta}}(\lambda)$ to be the minimizer of $Q(\boldsymbol{\beta}|\widehat{\mathbf{\Omega}}^{(1)})$, with $P_\lambda(\cdot)$ an oracle penalty function.
- Set $\widehat{\mathbf{\Omega}}(\lambda) = \widehat{\mathbf{\Omega}}\{\hat{\gamma}(\lambda)\}$ to be the minimizer of $Q\{\mathbf{\Omega}|\widehat{\boldsymbol{\beta}}(\lambda)\}$ with $P_\gamma(\cdot)$, where $\hat{\gamma}(\lambda)$ for given λ is selected using a selection criterion.

Choose "the best" estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Omega}}) = (\widehat{\boldsymbol{\beta}}(\hat{\lambda}), \widehat{\mathbf{\Omega}}(\hat{\gamma}(\hat{\lambda})))$ by estimating λ using a selection criterion, see Section 5.

The first stage provides simple consistent estimators of $\boldsymbol{\beta}$ and $\mathbf{\Omega}$, the second stage calculates their efficient counterparts. The proposed "profile" tuning parameter estimation at the second stage mimics the profile likelihood idea and is computationally efficient. For instance, when performing a double grid search, one has to iteratively estimate both $\boldsymbol{\beta}$ and $\mathbf{\Omega}$ for each combination of tuning parameters (λ, γ). Here, not all combinations of tuning parameter values need to be searched, and less iterations are required for each estimated combination of the tuning parameter values.

We can use the GLASSO procedure (Friedman, Hastie, and Tibshirani (2008)) to minimize $Q(\mathbf{\Omega}|\widehat{\boldsymbol{\beta}})$ if the L_1 penalty is used. Tuning parameter selection is discussed in Section 3.2. Standard errors of $\widehat{\boldsymbol{\beta}}$ can be estimated by a formula similar to (3.2). Rothman et al. (2010). In Section 4.2, we show that the two-stage procedure produces the oracle estimators of $(\boldsymbol{\beta}, \mathbf{\Omega})$.

3.2. Tuning parameter selection

We propose BIC criteria for selecting the tuning parameter when a working (possibly misspecified) precision matrix is used, or when $\boldsymbol{\beta}$ and $\mathbf{\Omega}$ are jointly estimated.

Denote by $\widehat{\boldsymbol{\beta}}(\lambda)$ the estimated regression coefficient given the tuning parameter λ. Let $\hat{s}$ be the number of estimated non-zero coefficients, $|\{j : \hat{\beta}_j \neq 0\}|$, and $k_n \to \infty$. When a known working precision matrix $\boldsymbol{\Lambda}$ is used for estimating $\boldsymbol{\beta}$ as in Section 3.1., the BIC for choosing λ is

$$BIC\{\widehat{\boldsymbol{\beta}}(\lambda)\} = \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_m(\widehat{\boldsymbol{\beta}}(\lambda))\boldsymbol{\Lambda}_m\right] + \hat{s}\frac{k_n}{n}. \tag{3.4}$$

This BIC is also appropriate when $\boldsymbol{\Lambda}$ is a consistent estimator of $\boldsymbol{\Omega}$, rather than a fixed matrix. This is relevant, for instance, when one first estimates $\boldsymbol{\Omega}$ based on some consistent estimator of $\boldsymbol{\beta}$ and then uses the single initial estimator of $\boldsymbol{\Omega}$, denoted by $\boldsymbol{\Lambda}$, to estimate $\boldsymbol{\beta}$ without iterations or a two dimensional grid search.

When $\boldsymbol{\beta}$ is jointly estimated with $\boldsymbol{\Omega}$ for a given λ, $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}\{\hat{\gamma}(\lambda)\}$ is a function of λ. This since, given $\widehat{\boldsymbol{\beta}}(\lambda)$, the tuning parameter γ is estimated as a function of $\widehat{\boldsymbol{\beta}}(\lambda)$ by minimizing with respect to γ(Gao et al. (2009))

$$BIC(\widehat{\boldsymbol{\Omega}}_m(\gamma)|\lambda) = -\log|\widehat{\boldsymbol{\Omega}}_m(\gamma)| + \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_m\{\widehat{\boldsymbol{\beta}}(\lambda)\}\widehat{\boldsymbol{\Omega}}_m(\gamma)\right] + \hat{t}\frac{\log(n)}{n}, \tag{3.5}$$

where $\widehat{\boldsymbol{\Sigma}}_m$ is the sample covariance matrix and $\hat{t}$ is the estimated number of non-zero entries in the upper off-diagonal matrix of $\widehat{\boldsymbol{\Omega}}_m(\gamma)$. We denote the resulting estimator as $\hat{\gamma}(\lambda)$ and $\widehat{\boldsymbol{\Omega}}_m(\lambda) = \widehat{\boldsymbol{\Omega}}_m\{\hat{\gamma}(\lambda)\}$. Hence we define the BIC for λ when Ω is estimated together with $\boldsymbol{\beta}$ as

$$BIC(\widehat{\boldsymbol{\beta}}(\lambda)) = -\log(|\widehat{\boldsymbol{\Omega}}_m(\lambda)|) + \operatorname{tr}\left[\widehat{\boldsymbol{\Omega}}_m(\lambda)\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}(\hat{\lambda})\right] + \hat{s}\frac{k_n}{n}. \tag{3.6}$$

The optimal tuning parameter λ is estimated by minimizing (3.6).

4. Asymptotic Results When $p < n$

We first study the properties of the penalized multivariate regression estimator of $\boldsymbol{\beta}$ assuming an arbitrary working precision matrix, when the penalty function satisfies the oracle properties (Fan and Peng (2004)). We then extend the results to joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. The proofs are given in the supplementary material.

4.1. Estimation of regression coefficients when the working precision matrix is given

We show that for an arbitrary working precision matrix, $\widehat{\boldsymbol{\beta}}$ is consistent and sparsistent, where "sparsistency" means that every true zero parameter is estimated as zero with probability tending to 1, uniformly over all parameters (Lam and Fan (2009)). We first provide results for the case where the number

of parameters is smaller than the sample size, $p/n \to 0$; in this case we also show that $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed and is most efficient when the working precision matrix $\boldsymbol{\Lambda}$ is correctly specified as $\boldsymbol{\Omega}_m$. We first outline the regularity conditions.

(C1) $n \to \infty$, $p = p_0 m$ may vary with n, and $p/n \to 0$.

(C2) There exists a positive constant R such that

$$0 < 1/R < \begin{matrix} \lambda_{min}(\boldsymbol{\Lambda}_m), \lambda_{min}(\boldsymbol{\Omega}_m), \lambda_{min}(n^{-1}\mathbf{X}^T\mathbf{X}) \\ \lambda_{max}(\boldsymbol{\Lambda}_m), \lambda_{max}(\boldsymbol{\Omega}_m), \lambda_{max}(n^{-1}\mathbf{X}^T\mathbf{X}) \end{matrix} < R < \infty,$$

where $\lambda_{min}(\boldsymbol{B})$ and $\lambda_{max}(\boldsymbol{B})$ are the smallest and largest eigenvalues of a matrix $\boldsymbol{B}$.

(C3) If $\rho = \rho_n = \min\{|\beta_j|;\ j \in A\}$. $\rho/\sqrt{p/n} \to \infty$.

(C4) $\max_{1 \le i \le n} n^{-1}||\boldsymbol{x}_i \boldsymbol{x}_i^T||_2 \to 0$, where $||\boldsymbol{x}_i \boldsymbol{x}_i^T||_2 = ||\boldsymbol{x}_i||_2^2$.

(C5) There exists a δ such that $E(\epsilon_{ij}^{2+\delta}) < \infty$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

(C6) The function P_λ is concave on $[0, \infty)$ and differentiable on $(0, \infty)$, with $P_\lambda(0) = 0$, $P_\lambda(\theta) = P_\lambda(-\theta)$ and $\lim_{\theta \to \infty} P_\lambda(\theta) \le 1/n$.

(C7) If $r_n/\sqrt{p/n} \to \infty$, then $P'_\lambda(r_n) = o(1/\sqrt{np})$.

(C8) Let k_n be such that $p/k_n^{(2+\delta)/2} \to 0$. If $r_n = O(\sqrt{p/n})$, then $\lim_{n \to \infty} (\sqrt{n/\max(p, k_n)}) P'_\lambda(r_n) \to \infty$.

Conditions (C1)−(C5) are related to the likelihood function. Conditions (C6)−(C8) are related to the penalty function. The latter two conditions are specified so that the behavior of the derivative of the penalty function that is "close to zero" and "far from zero" in some rate sense behaves according to the rate of convergence of the estimators. These requirements are guaranteed to hold for the oracle penalty SELO, for instance, for some specifications of sequences of the tuning parameters.

Specifically, condition (C1) bounds the rate of the number of covariates p. Condition (C2) guarantees the stability of the estimator. Condition (C3) sets a required bound on the smallest parameter in the model and permits model selection consistency. Condition (C4) gives a very weak bound on the covariates. Notice that it could equivalently be written as $\max_{1 \le i \le n} n^{-1} \boldsymbol{x}_i^T \boldsymbol{x}_i \to 0$ which is not restrictive as covariates, for instance, are often naturally bounded. The last requirement on the likelihood function, (C5) is important for asymptotic normality of the regression coefficient estimators.

Under these conditions, the estimator $\widehat{\boldsymbol{\beta}}$ of the regression parameters obtained by minimizing $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda})$ in (3.1) has the oracle properties.

Theorem 1. *Suppose* (C1)−(C8) *hold. More specifically, if* $\boldsymbol{\beta}^*$ *is the true vector of regression parameters, then the following hold.*

(a) (*Consistency*) *For every* $c > 0$, *there exists a positive constant* M *such that*

$$\liminf_{n\to\infty} P\Big[\text{There exists a local minimum } \widehat{\boldsymbol{\beta}} \text{ of } Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m) \text{ such that } ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| < M\sqrt{\frac{p}{n}}\Big] > 1 - c.$$

Further, a global minimizer of $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m)$ *is* $\sqrt{p/n}$ *consistent for* $\boldsymbol{\beta}^*$.

(b) (*Sparsistency*) *For every fixed* $M > 0$,

$$\lim_{n\to\infty} P\Big[\Big\{\widehat{\boldsymbol{\beta}} \text{ is a local minimum of } L(\boldsymbol{\beta}) \text{ and } ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| \le M\sqrt{\frac{p}{n}}\Big\} \cap \Big\{\widehat{\boldsymbol{\beta}}_{A^c} \ne 0\Big\}\Big] = 0.$$

(c) (*Asymptotic normality*) *Let* $s = |A| = |\{j; \beta_j^* \ne 0\}$. *Let* $q \in \mathbb{N}$ *be fixed and* $\boldsymbol{B} = \boldsymbol{B}_n$ *be a sequence of* $q \times s$ *matrices such that* $\boldsymbol{B}\boldsymbol{B}^T \to \boldsymbol{G}$, *for some* $q \times q$ *symmetric matrix* $\boldsymbol{G}$. *There exists a sequence of local minima,* $\widehat{\boldsymbol{\beta}}$, *of* $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m)$ *such that* $\lim_{n\to\infty} P(\{j;\ \hat{\beta}_j \ne 0\} = A) = 1$. *And*

$$\boldsymbol{B}\check{\boldsymbol{\Upsilon}}^{-1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \xrightarrow{D} N(0, \boldsymbol{G}),$$

where $\check{\boldsymbol{\Upsilon}} = (\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_A^T\boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\mathbf{X}_A(\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}$. *If* $\boldsymbol{\Lambda} = \boldsymbol{\Omega}$, *then* $\boldsymbol{\Upsilon} = (\mathbf{X}_A^T\boldsymbol{\Omega}\mathbf{X}_A)^{-1}$, *and* $\widehat{\boldsymbol{\beta}}$ *is the most efficient estimator.*

4.2. Joint estimation of regression coefficients and precision matrix

We show consistency and sparsistency of the two-stage estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}})$. We first state additional regularity conditions required for estimation of the precision matrix. Here $P_\gamma(\cdot)$ is the penalty function used for the entries $\omega_{ij}, i \ne j = 1, \ldots, m$ of the precision matrix. Let $B = \{(i,j)|\ \omega_{ij} \ne 0, i < j\}$, the set of true non-zero entries in $\boldsymbol{\Omega}$, and $t = |B|$. Let $\tau = \tau_n = \min\{|\omega_{ij}| : (i,j) \in B\}$. Assume that $P_\gamma(\cdot)$ is an oracle penalty function that satisfies condition (C6), and the following.

(C9) $pm, m^2/n \to 0$.

(C10) $\sup_{j,k} E(\epsilon_{ij}\epsilon_{ik})^2 < \infty$.

(C11) If $r_n/\sqrt{(m+p)m/n} \to \infty$, then $mP_\gamma'(r_n) = O(1)$.

(C12) If $r_n = O(\sqrt{(m+p)m/n})$, then $\lim_{n\to\infty} \sqrt{(m+p)m/n}P_\gamma'(r_n) = \infty$.

Conditions (C11) and (C12) are satisfied under a proper selection of the sequence $\gamma_n = \gamma$. Let $||\mathbf{A}||_F$ be the Frobenius norm of the matrix $\mathbf{A}$.

Lemma 1. *Suppose that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_n$ is a sequence of estimators such that $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_P(\sqrt{p/n})$, and that* (C1)–(C8) *hold.*

(a) *If*

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^T, \text{ then } ||\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_m||_F = O_P\left\{\sqrt{(m+p)\frac{m}{n}}\right\}.$$

(b) *If* (C9)–(C12) *also hold. There exists a $\sqrt{(m+p)m/n}$ consistent local minimizer, $\widehat{\boldsymbol{\Omega}}_m$ of $Q(\boldsymbol{\Omega}_m|\widehat{\boldsymbol{\Sigma}}_m)$.*

(c) *Any local minimizer of $Q(\boldsymbol{\Omega}|\widehat{\boldsymbol{\Sigma}})$ satisfies $||\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}||_F^2 = O_P((m+p)m/n)$, and $\hat{\omega}_{jj'} = 0$ with probability 1 if $\omega_{jj'} = 0$, for all $(i,j) \in B$.*

Our results show that estimation of regression coefficients $\boldsymbol{\beta}$ increases the rate of convergence of the precision matrix estimator compared to that in non-regression settings (assuming mean 0 for $\boldsymbol{y}$). We make a more robust assumption on the error distribution, and thus the estimator $\widehat{\boldsymbol{\Omega}}$ requires a smaller rate on the number of outcomes that could otherwise be achieved.

Theorem 2. *Let $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}})$ be the estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$ obtained by the two-stage procedure.*

(i) *If* (C1)–(C8) *hold, then $\widehat{\boldsymbol{\beta}}$ is $\sqrt{p/n}$-consistent for $\boldsymbol{\beta}$, $\hat{\beta}_j = 0$ with probability tending to 1 for all j such that $\beta_j = 0$, also, $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed with parameters as in Theorem* 1(c).

(ii) *If also* (C9)–(C12) *hold, then $\widehat{\boldsymbol{\Omega}}$ is $\sqrt{(m+p)m/n}$- consistent for $\boldsymbol{\Omega}$, and $\hat{\omega}_{ij} = 0$ with probability tending to 1 for all i,j such that $\omega_{ij} = 0$. If $\boldsymbol{B}$ is a sequence of matrices as in Theorem* 1, *then $\boldsymbol{B}\boldsymbol{\Upsilon}^{-1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \xrightarrow{D} N(0, \boldsymbol{G})$, where $\boldsymbol{\Upsilon} = (\mathbf{X}_A^T\boldsymbol{\Omega}\mathbf{X}_A)^{-1}$.*

4.3. Model selection consistency of the BIC criteria for regression parameters

We show that, under regularity conditions, minimization of the BIC criterion leads to selection of the true model for $\boldsymbol{\beta}$ asymptotically, either when $\boldsymbol{\Lambda}_m$ is an arbitrary working precision matrix or when $\boldsymbol{\Lambda}_m = \widehat{\boldsymbol{\Omega}}_m$ is jointly consistently estimated. Let $x_n = \Theta(y_n)$ denote sequences satisfying $x_n/y_n \to \infty$.

Theorem 3.

(a) *Suppose* (C1)−(C8) *hold, with* $\beta_j \neq 0 \Leftrightarrow |\beta_j| = \Theta(\sqrt{\max\{k_n, p\}/n})$. *Let* $\widehat{\boldsymbol{\beta}}_{\hat{A}}$ *be a minimizer of* $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda})$ *for any tuning parameter value* λ *such that the non-zero entries in* $\widehat{\boldsymbol{\beta}}_{\hat{A}}$ *correspond to a model* $\hat{A} \neq A$. *Then if* $\boldsymbol{\Lambda}$ *is a fixed positive definite matrix,*

$$P\left\{\sup_{\hat{A}\neq A, \hat{A}\in\mathcal{A}} \left(BIC(\widehat{\boldsymbol{\beta}}_A) - BIC(\widehat{\boldsymbol{\beta}}_{\hat{A}})\right) < 0\right\} \to 1 \text{ as } n \to \infty.$$

(b) *If* $\boldsymbol{\Lambda}_m$ *is a consistent estimator of* $\boldsymbol{\Omega}_m$, *obtained using any* $\sqrt{p/n}$*-consistent estimator of the mean* $\mathbf{X}\widehat{\boldsymbol{\beta}}$ *of* $\mathbf{Y}$.

$$P\left\{\sup_{\hat{A}\neq A, \hat{A}\in\mathcal{A}} \left(BIC(\widehat{\boldsymbol{\beta}}_A) - BIC(\widehat{\boldsymbol{\beta}}_{\hat{A}})\right) < 0\right\} \to 1 \text{ as } n \to \infty.$$

This result shows that the BIC in (3.4) gives an estimator of $\boldsymbol{\beta}$ that is consistent for model selection when an arbitrary working precision matrix $\boldsymbol{\Lambda}$ is used, and in case $\boldsymbol{\Lambda}$ is a consistent estimator of the precision matrix.

5. Asymptotic Results When $p > n$

The use of a working covariance matrix leads to an estimator of the regression coefficient vector that is consistent under suitable regularity conditions. We modify the conditions for the rate that $p/n \to 0$ to the following.

(C1′) With $m < n < p$, $\log(p)/n \to 0$. The true model size $s = |A|$ satisfies $s < n$ and $s/\log(p) \to 0$.

(C2′) The eigenvalues of the positive definite matrices $(1/n)\mathbf{X}\mathbf{X}^T$, $(1/n)\mathbf{X}_A^T\mathbf{X}_A$ satisfy

$$0 < R^{-1} < \lambda_{\min}\left(\frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A\right), \lambda_{\min}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A\right),$$
$$\lambda_{\max}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right) < R < \infty.$$

(C3′) $\min_{j:\beta_j\neq 0} \frac{\beta_j}{\sqrt{\log(p)/n}} \to \infty$.

(C7′) If $\lim_n \frac{r_n}{\sqrt{\log(p)/n}} = \infty$, then $n\sqrt{\log(p)/n}P'_\lambda(r_n) = o(1)$.

(C8′) If $\lim_n \frac{r_n}{\sqrt{\log(p)/n}} \leq c$, then $P'_\lambda(r_n)/m \to \infty$.

(C10′) The errors are normal, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Theorem 4. *Suppose that* $\boldsymbol{\Lambda}_m$ *is given, and that conditions* (C1')−(C8') *hold. with the above modifications. Let* $\boldsymbol{\beta}^*$ *be the true vector of regression parameters.*

(a) (*Consistency*) *For every* $c > 0$, *there exists a positive constant* M *such that*

$$\liminf_{n\to\infty} P\Bigg[\textit{There exists a local minimum } \widehat{\boldsymbol{\beta}} \textit{ of } Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m) \textit{ such that } ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| < M\sqrt{\frac{\log(p)}{n}}\Bigg] > 1 - c.$$

(b) (*Sparsistency*) *For every fixed* $M > 0$,

$$\lim_{n\to\infty} P\Bigg[\Bigg\{\widehat{\boldsymbol{\beta}} \textit{ is a local minimum of } L(\boldsymbol{\beta}) \textit{ and } ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| \le M\sqrt{\frac{\log(p)}{n}}\Bigg\} \cap \Big\{\widehat{\boldsymbol{\beta}}_{A^c} \neq 0\Big\}\Bigg] = 0.$$

If the number of outcomes satisfies $m^2/n \to 0$, then these asymptotic properties readily extend to joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. Let (C9)−(C12) be modified, with p replaced by $\log(p)$, to get conditions (C9′)−(C12′).

Corollary 1. *Let* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}})$ *be the estimators of* $(\boldsymbol{\beta}, \boldsymbol{\Omega})$ *obtained by the two-stage procedure.*

(i) *If* (C1′)-(C8′) *hold, then* $\widehat{\boldsymbol{\beta}}$ *is a* $\sqrt{\log(p)/n}$*-consistent for* $\boldsymbol{\beta}$, *and* $\hat{\beta}_j = 0$ *with probability tending to 1 for all* j *such that* $\beta_j = 0$.

(ii) *If* (C1′)−(C12′) *hold, then* $\widehat{\boldsymbol{\Omega}}$ *is* $\sqrt{(m + \log(p))m/n}$*-consistent for* $\boldsymbol{\Omega}$, *and* $\hat{\omega}_{ij} = 0$ *with probability tending to 1 for all* i, j *such that* $\omega_{ij} = 0$.

6. Simulation Studies

We present simulation results as evaluation of the finite sample performance of the proposed methods. Additional simulation results are provided in the supplementary material.

6.1. Small-p simulations

We consider two types of correlation structures between outcomes: autoregressive (AR) and exchangeable (EX) outcome correlations. In both scenarios the correlation parameter was $\rho_y = 0.5$, and the variance was 9. The precision matrix is thus sparse in the AR case (tridiagonal) and dense in the EX case. In each of the simulations we generated five outcomes and five covariates, so $p = 25$. The parameter vector $\boldsymbol{\beta}_{25\times 1}$ was generated such that three entries were non-zeros, and their value were 3, 2, and 1.5 (as in Tibshirani (1996)). The predictor $\mathbf{X}$ was simulated by assuming its covariance to be autoregressive with $\rho_x = 0.7$.

For each scenario, $n = 50, 100$, and 200 simulations were run for each sample size. We estimated $\boldsymbol{\beta}$ using the Lasso, adaptive Lasso, SCAD (with $a = 3.7$) and SELO (with $\tau = 0.01$). The precision matrix $\boldsymbol{\Omega}$ was estimated using GLASSO with the Lasso penalty. (The Lasso penalty does not satisfy (C11)−(C12).

Tuning parameters were selected either by BIC or by a validation data set. For the latter, we minimized prediction error. Thus, for $\mathbf{X}_{valid}$ and $\mathbf{Y}_{valid}$ independent validation data sets, we estimated λ by minimizing

$$||\mathbf{Y}_{valid} - \mathbf{X}_{valid}\widehat{\boldsymbol{\beta}}(\lambda)||_2. \tag{6.1}$$

Given $\boldsymbol{\beta}(\lambda)$, to select the tuning parameter γ for estimating $\boldsymbol{\Omega}$, we chose the γ that maximizes the loglikelihood of the validation data as

$$\log\left|\widehat{\boldsymbol{\Omega}}(\gamma)\right| - \mathrm{tr}\{\widehat{\boldsymbol{\Omega}}(\gamma)\widehat{\boldsymbol{\Sigma}}_{valid}\}, \tag{6.2}$$

where $\widehat{\boldsymbol{\Omega}}(\gamma) = \widehat{\boldsymbol{\Sigma}}^{-1}(\gamma)$ is estimated from the training data, and $\widehat{\boldsymbol{\Sigma}}_{valid}$ is the sample covariance matrix estimated from the validation data set. The same validation data set was used to calculate (6.1) and (6.2). For tuning parameters selection using a validation data set, we used twice the number of observations so that, for $n = 100$, we used 100 observations for parameters estimation and additional 100 observations for validation. For the BIC, we used only 100 observations.

To evaluate the performance of variable selection procedures, we summarize the results with averages of the following measures across the 200 simulations. Model size: $|\{\hat{\beta}_{ij} \neq 0\}|$; the true model size in all simulations was 3. True model chosen: $\mathbf{1}(\{i, j : \hat{\beta}_{ij} \neq 0\} = \mathcal{A})$. False positives: $|\{\hat{\beta}_{ij} \neq 0, \beta_{ij} = 0\}|$. False negatives: $|\{\hat{\beta}_{ij} = 0, \beta_{ij} \neq 0\}|$. Model error (Yuan and Lin (2007)): $\mathrm{tr}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}_{XX}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$. Prediction error: for $\mathbf{X}_{new}, \mathbf{Y}_{new}$ validation data sets, $||\mathbf{Y}_{new} - \mathbf{X}_{new}\widehat{\boldsymbol{\beta}}||_2$.

For comparison, we also used the MRCE algorithm (Rothman et al. (2008)) to estimate the regression parameters. This algorithm performs a double grid search over the tuning parameter values for the regression parameters and the precision matrix using the Lasso penalty. To select the best combination of tuning parameters, cross validation was used. As such, the results of the MRCE are most appropriately compared to the results of the two-stage algorithm implemented with the Lasso penalty. Since cross validation was used for tuning parameter selection with MRCE, the total sample size used in each simulation was identical to that used with the two-stage algorithm with BIC, while the two-stage algorithm with validation used twice the sample size.

The simulation results in Table 1 show that selecting the tuning parameter by minimizing prediction error using independent validation data yields larger

Table 1. Simulation results over 200 simulations using the two-stage joint estimation procedure with the penalties LASSO, adaptive LASSO, SCAD, and SELO. The precision matrix was estimated using GLASSO with the LASSO penalty. Sample sizes were $n = 50, 100$. The number of outcomes was five and the number of covariates was five. The correlation structures considered were: AR(1) and exchangeable (EX). The tuning parameters was selected either using a validation data set with the same sample size, or by BIC. The average selected model size, percentage of times the true model being selected, average number of false positives, average number of false negatives, average model error and prediction error are reported. The results are compared to the MRCE algorithm.

Simulation	Method	size	True model	false pos	false neg	ME	pred err
AR cov, $n=50$	LASSO - BIC	5.31	0.14	2.39	0.08	1.70	9.53
	Adaptive LASSO - BIC	3.88	0.32	1.26	0.38	1.59	9.60
	SCAD - BIC	4.22	0.19	1.71	0.48	2.06	9.71
	SELO - BIC	3.82	0.38	1.14	0.32	1.68	9.66
	LASSO - validation	9.13	0.04	6.17	0.04	1.41	9.51
	Adaptive Lasso - validation	6.46	0.08	3.69	0.24	1.39	9.56
	SCAD - validation	5.18	0.10	2.54	0.36	1.52	9.62
	SELO - validation	3.75	0.46	1.10	0.35	1.39	9.60
	MRCE	8.35	0.01	5.41	0.06	1.71	11.59
AR cov, $n=100$	Lasso - BIC	4.88	0.19	1.88	0.00	0.88	9.42
	Adaptive LASSO - BIC	3.66	0.48	0.80	0.14	0.63	9.45
	SCAD - BIC	3.54	0.60	0.73	0.18	0.59	9.46
	SELO - BIC	3.65	0.57	0.76	0.10	0.58	9.46
	LASSO - validation	9.80	0.02	6.80	0.00	0.73	9.42
	Adaptive LASSO - validation	6.45	0.14	3.50	0.06	0.60	9.44
	SCAD - validation	5.22	0.30	2.31	0.09	0.51	9.45
	SELO - validation	4.00	0.65	1.09	0.10	0.50	9.44
	MRCE	8.74	0.02	5.74	0.00	0.82	11.41
EX cov, $n=50$	LASSO - BIC	5.25	0.17	2.29	0.05	1.54	9.36
	Adaptive LASSO - BIC	4.13	0.30	1.43	0.30	1.37	9.28
	SCAD - BIC	3.81	0.30	1.21	0.40	1.51	9.36
	SELO - BIC	3.79	0.44	1.04	0.25	1.36	9.27
	LASSO - validation	9.53	0.02	6.54	0.02	1.33	9.29
	Adaptive LASSO - validation	6.54	0.10	3.73	0.19	1.30	9.25
	SCAD - validation	5.34	0.16	2.63	0.28	1.26	9.25
	SELO - validation	3.75	0.50	1.04	0.30	1.15	9.23
	MRCE	8.26	0.01	5.29	0.03	1.39	10.98
EX cov, $n=100$	LASSO - BIC	4.88	0.20	1.88	0.00	0.82	9.11
	Adaptive LASSO - BIC	3.92	0.50	1.01	0.10	0.54	9.13
	SCAD - BIC	3.42	0.66	0.54	0.12	0.44	9.11
	SELO - BIC	3.54	0.63	0.58	0.04	0.42	9.10
	LASSO - validation	10.29	0.02	7.29	0.00	0.68	9.11
	Adaptive LASSO - validation	6.80	0.18	3.83	0.03	0.54	9.12
	SCAD - validation	5.16	0.36	2.23	0.07	0.45	9.12
	SELO - validation	4.12	0.71	1.18	0.06	0.40	9.11
	MRCE	8.91	0.01	5.92	0.00	0.78	10.90

Table 2. Median estimated SEs, denoted by SE_EST for the four simulation configurations described in Table 6.1 over 200 simulations. The SE estimators were calculated using the sandwich formula. Empirical SEs calculated using the median absolute deviation of the parameter estimates (SE_EMP) are provided in parenthesis.

Simulation	Method	$\beta_1 = 3$	$\beta_2 = 1.5$,	$\beta_3 = 2$,
		SE_EST(SE_EMP)	SE_EST(SE_EMP)	SE_EST(SE_EMP)
AR cov, $n=50$	LASSO - BIC	0.39 (0.43)	0.33 (0.37)	0.35 (0.34)
	Adaptive LASSO - BIC	0.41 (0.55)	0.33 (0.71)	0.35 (0.41)
	SCAD - BIC	0.48 (0.62)	0.44 (1.05)	0.38 (0.37)
	SELO - BIC	0.5 (0.49)	0.49 (0.59)	0.38 (0.36)
	LASSO - validation	0.44 (0.38)	0.43 (0.36)	0.37 (0.35)
	Adaptive LASSO - validation	0.44 (0.49)	0.43 (0.6)	0.37 (0.37)
	SCAD - validation	0.46 (0.59)	0.46 (0.93)	0.38 (0.38)
	SELO - validation	0.46 (0.52)	0.45 (0.69)	0.35 (0.3)
AR cov, $n=100$	LASSO - BIC	0.32 (0.25)	0.29 (0.25)	0.27 (0.24)
	Adaptive LASSO - BIC	0.33 (0.34)	0.3 (0.33)	0.27 (0.26)
	SCAD - BIC	0.37 (0.32)	0.38 (0.36)	0.27 (0.19)
	SELO - BIC	0.38 (0.27)	0.38 (0.31)	0.28 (0.2)
	LASSO - validation	0.34 (0.26)	0.34 (0.26)	0.29 (0.24)
	Adaptive LASSO - validation	0.34 (0.3)	0.34 (0.34)	0.28 (0.25)
	SCAD - validation	0.35 (0.31)	0.36 (0.34)	0.26 (0.19)
	SELO - validation	0.35 (0.29)	0.35 (0.32)	0.26 (0.19)
EX cov, $n=50$	LASSO - BIC	0.35 (0.37)	0.3 (0.36)	0.32 (0.33)
	Adaptive LASSO - BIC	0.36 (0.48)	0.28 (0.61)	0.32 (0.36)
	SCAD - BIC	0.42 (0.52)	0.42 (0.8)	0.33 (0.36)
	SELO - BIC	0.43 (0.43)	0.43 (0.48)	0.34 (0.32)
	LASSO - validation	0.4 (0.36)	0.39 (0.33)	0.34 (0.31)
	Adaptive LASSO - validation	0.4 (0.46)	0.39 (0.53)	0.33 (0.36)
	SCAD - validation	0.42 (0.53)	0.42 (0.75)	0.33 (0.33)
	SELO - validation	0.41 (0.47)	0.4 (0.6)	0.31 (0.29)
EX cov, $n=100$	LASSO - BIC	0.27 (0.23)	0.26 (0.21)	0.25 (0.21)
	Adaptive LASSO - BIC	0.29 (0.29)	0.27 (0.32)	0.24 (0.22)
	SCAD - BIC	0.33 (0.26)	0.33 (0.27)	0.24 (0.19)
	SELO - BIC	0.33 (0.23)	0.33 (0.24)	0.25 (0.2)
	LASSO - validation	0.3 (0.21)	0.31 (0.23)	0.27 (0.21)
	Adaptive LASSO - validation	0.31 (0.27)	0.31 (0.3)	0.25 (0.21)
	SCAD - validation	0.31 (0.27)	0.32 (0.28)	0.23 (0.18)
	SELO - validation	0.31 (0.24)	0.32 (0.24)	0.23 (0.18)

models than that selected using BIC. Similarly, model error was usually better when using validation data for model selection. However, BIC selected a higher proportion of true models. Lasso tended to select larger models, and SCAD and SELO selected the true model more often. In terms of variable selection, MRCE performed better than the two-stage approach using the Lasso penalty when the tuning parameter was selected using validation, but worse when the

tuning parameter was selected using BIC. The prediction error with MRCE was worse than when using Lasso with validation.

In addition to estimating $\mathbf{\Omega}$ sparsely using GLASSO, we also examined the effect of other estimators of the precision matrix on the estimator of $\boldsymbol{\beta}$. We compared the estimator of $\mathbf{\Omega}$ using ridge, and the shrinkage estimator of Schäfer and Strimmer (2005). The GLASSO yielded the most stable results compared to other regularized methods of $\mathbf{\Omega}$. In addition, we also compared the GLASSO to a parametric estimator of the precision matrix, when the true covariance structure was assumed, and when independent correlation structure was assumed. Results for these scenarios are provided in the supplementary material. When using a parametric estimator of the precision matrix, model selection results were comparable to those using GLASSO in both EX (dense precision matrix) and AR (sparse precision matrix) settings. When assuming working independence, the estimated models were slightly larger, with the true model selected fewer times on average.

Table 2 compares the empirical SEs with the estimated SEs using the sandwich formula for the non-zero estimated regression parameters. The results show that the standard error estimates had mixed performance. When $n = 100$, usually they were slightly larger than the empirical SEs and, in general, closer to the empirical ones when the β coefficients were larger. When $n = 50$, the pattern was less uniform, with especially larger differences between the estimated and empirical SEs when $\beta = 1.5$, suggesting when β is relatively small, the SE estimates tend to underestimate the true variability.

6.2. Large-p simulations

We adapt the small-p simulations to large-p scenarios. In all simulations we had $n = 50$ observations. The number of covariates and the number of outcomes in $\{5, 20\} \times \{5, 20\}$. With $(p_0, m = 5)$ this is a "small-p" (25 coefficients) scenario; two scenarios were with 100 coefficients, and one had 400 coefficients to estimate. We simulated AR and EX outcome correlations, each with $\rho_y = \text{cor}(Y_k, Y_{k-1}) = 0.5$ and $\text{var}(Y_k) = 3$. The covariates $\mathbf{X}$ had an autoregressive covariance structure, with $\text{cor}(X_k, X_l) = 0.3$, $k, l = 1, \ldots, p$. We conducted 200 simulations for each scenario, and applied the penalty functions Lasso, adaptive Lasso (with the initial estimator of $\boldsymbol{\beta}$ estimated as the weight using ridge regression with the General Cross Validation criterion (GCV)), SCAD (with $a = 3.7$), and SELO (with $\tau = 0.01$).

Tuning parameters were selected using data validation, with minimization of prediction error. Thus, 50 additional observations were sampled from the same distribution and used for tuning parameter selection. Here BIC was not used for tuning parameter selection. The true model consisted of three non-zero parameter values, $(3, 1.5, 2)$. Table 3 provides the proportion of the true model

Table 3. Large-p simulation results averaged over 200 simulations using the two-stage joint estimation procedure with LASSO, adaptive LASSO, SCAD, and SELO, with the precision matrix estimated using GLASSO with the Lasso penalty. The sample size was $n = 50$. The number of covariates and the number of outcomes were, independently, 5 and 20. The true model size were three. EX and AR correlations were considered. The tuning parameters were selected using an independent validation data set with the same sample size by minimizing the prediction error. The percentage of times the true model being chosen (mean T), the number of variables that were selected as false positives (mean FP), the average number of variables that were selected as false negatives (mean FN), and the average model error (mean ME) are reported.

penalty	mean T	mean FP	mean FN	mean ME
AR cov				
$p_0 = 5,\ m = 5$				
LASSO	0.01	8.65	0.00	0.57
Adaptive LASSO	0.00	16.05	0.05	1.39
SCAD	0.31	3.41	0.00	0.29
SELO	0.67	0.76	0.00	0.21
$p_0 = 20,\ m = 5$				
LASSO	0.00	13.83	0.00	0.80
Adaptive LASSO	0.00	26.68	0.16	2.21
SCAD	0.14	6.19	0.00	0.30
SELO	0.72	0.51	0.00	0.24
$p_0 = 5,\ m = 20$				
LASSO	0.00	18.36	0.00	1.03
Adaptive LASSO	0.00	45.55	0.11	3.03
SCAD	0.07	7.21	0.00	0.39
SELO	0.56	1.38	0.00	0.37
$p_0 = 20,\ m = 20$				
LASSO	0.00	23.81	0.00	1.37
Adaptive LASSO	0.00	41.94	0.48	4.12
SCAD	0.03	14.63	0.00	0.58
SELO	0.65	0.87	0.01	0.36

penalty	mean T	mean FP	mean FN	mean ME
EX cov				
$p_0 = 5,\ m = 5$				
LASSO	0.01	9.01	0.00	0.52
Adaptive LASSO	0.00	15.63	0.06	1.24
SCAD	0.29	3.42	0.00	0.25
SELO	0.66	0.86	0.00	0.19
$p_0 = 20,\ m = 5$				
LASSO	0.00	14.12	0.00	0.75
Adaptive LASSO	0.00	26.56	0.19	2.19
SCAD	0.16	6.24	0.00	0.28
SELO	0.70	0.61	0.00	0.23
$p_0 = 5,\ m = 20$				
LASSO	0.00	17.41	0.00	0.75
Adaptive LASSO	0.00	44.29	0.06	1.97
SCAD	0.12	6.43	0.00	0.28
SELO	0.62	1.52	0.00	0.27
$p_0 = 20,\ m = 20$				
LASSO	0.00	22.50	0.00	0.98
Adaptive LASSO	0.00	52.02	0.21	2.83
SCAD	0.09	12.53	0.00	0.36
SELO	0.61	1.31	0.00	0.29

chosen, the average numbers of false positives and false negatives, and the average squared model error defined as $\mathrm{tr}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{\Sigma}_{XX} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$.

The size of the precision matrix was larger when $m = 20$ then when $m = 5$. Therefore, the simulation results were better when $m = 5, p_0 = 20$ than when $m = 20, p_0 = 5$, with a higher proportion of true models chosen and lower numbers of false positives.

The outcome correlations lead to roughly similar results. Comparing the performance of different penalties, SELO had the best results, SCAD performed the second best with more false positive detections, and the Lasso and adaptive Lasso had considerably higher numbers of false positives. Adaptive Lasso had the worst results, with the highest false positive rate and model error.

Results from additional simulations, in which parametric and working independent correlation structures were used, are provided in the supplementary material. The results show similar patterns to those in Table 3.

7. Analysis of the Type 2 Diabetes Data

The type 2 diabetes data set (Mootha (2003)) consists of gene expression profiles of 43 males of similar age with three levels of Glucose Tolerance: Normal (NGT, 17 subjects), Impaired (IGT, 9 subjects) and type 2 Diabetes Mellitus (DM2, 17 subjects). The authors provided a list of gene sets, and performed gene set analysis by testing for the global effect of glucose tolerance levels on each gene set. As many genes in a gene set are likely not affected by glucose levels, we are interested in selecting a subset of genes in a gene set that are associated with glucose tolerance levels. We focused on the porphyrin and chlorophyll metabolism pathway, which has gene expressions of 35 probes of genes. After removing unexpressed probes, 18 of them were used for analysis. We treated the expresses probes as outcomes whereas the covariates were two dummy variables for glucose tolerance levels, with NGT as the baseline. We considered the model

$$y_{ij} = \beta_{0j} + \beta_{1j}\mathbf{1}\mathrm{IGT}_i + \beta_{2j}\mathbf{1}\mathrm{DM}_i + e_{ij},$$

where y_{ij} is the j^{th} outcome (probe expression) of subject i, $\mathrm{IGT}_i = 1$ if a subject has impaired glucose tolerance and 0 otherwise, and DM_i takes value 1 if a subject has DM2 and 0 otherwise. Hence $m = 18$, $p = 18 \times 3 = 54$ and $n = 43$. We applied the proposed two-stage procedure with the penalties Lasso, adaptive Lasso, SCAD, and SELO for estimation of $\boldsymbol{\beta}$. The 18×18 precision matrix was estimated using the Lasso penalty via the GLASSO algorithm. The tuning parameters were selected using BIC for both regression coefficients and precision matrix estimation.

Table 4 presents the results of the four penalized regression models. Estimates of the β_{1j} and β_{2j}, the effects of glucose tolerance groups IGT and DM2 on probe expression relative to the baseline NGT group, are presented. The P-values of non-zero estimated coefficients are given in parentheses. The baseline effect (intercept estimates) are omitted. There are more non-zero estimates for the coefficients of DM than those for the coefficients of IGT. This makes sense biologically, as IGT is more similar to NGT than DM.

With all penalties, DM was found to be a significant predictor of the expressions of EPRS(1), HMBS, and ADH6, IGT is a significant predictor for HMBS and CP. IGT was also a significant predictor of BLVRB in the models selected by SELO, Lasso, SCAD, but not by adaptive Lasso. DM was found to be significantly associated with HMOX2 using SELO. The gene HMBS was found to be associated with with IGT and DM.

Lasso produces a large number of non-zero parameter estimates compared to the oracle penalties. The Lasso parameter estimates are shrunk more than those estimated by the oracle penalties and the P-values are relatively high. The SELO penalty produced the smallest number of non-zero parameter estimates.

The data set and code used in this analysis are provided in the journal website.

Table 4. Analysis results of the diabetes data set ($n = 43$) with the porhphyrin and chlorophyl metabolism pathway, which has 18 genes ($m = 18$). Variable selection was performed using the penalized multivariate regression regression with the two-stage joint estimation procedure with the penalties LASSO, adaptive LASSO, SCAD and SELO. The baseline expression values (intercepts), representing the mean expression levels of the normal glucose tolerance group are not presented. IGT and DM correspond to the mean expression differences between subjects with Impaired Glucose Tolerance (IGT) and subjects with type II diabetes (DM) compared the normal group. The number of regression coefficients was $p = 18 \times 3 = 54$. The tuning parameter was estimated using the BIC. The P-values are in parentheses.

Gene (probe)	Lasso		Adaptive Lasso		Scad		SELO	
	IGT	DM	IGT	DM	NGT	DM	IGT	DM
EPRS (1)	0.01 (0.88)	0.62 (0.01)	0	0.69 (0.005)	0	0.73 (0.004)	0	0.79 (0.002)
EPRS (2)	0.27 (0.19)	0.15 (0.41)	0.29 (0.18)	0.12 (0.49)	0.11 (0.49)	0 (0.83)	0	0
EPRS (3)	0	0.06 (0.53)	0	$<$ 0.01 (0.88)	0	0.02 (0.75)	0	0
BLVRB	0.35 (0.007)	-0.07 (0.61)	0.36 (0.478)	-0.03 (0.98)	0.44 (0.003)	0	0.5 (0.001)	0
GUSB	-0.02 (0.86)	-0.13 (0.5)	0	-0.11 (0.82)	0	-0.07 (0.67)	0	0
UROS	-0.08 (0.56)	0	-0.06 (0.84)	0	-0.02 (0.76)	0	0	0
HMBS	-0.16 (0.15)	-0.23 (0.05)	-0.18 (0.17)	-0.26 (0.05)	-0.13 (0.27)	-0.26 (0.04)	-0.25 (0.03)	-0.32 (0.006)
FECH	-0.03 (0.63)	0	0	0	0	0	0	0
HMOX1	0.03 (0.76)	0	0	0	0	0	0	0
HCCS (1)	0	-0.03 (0.69)	0	0	0	0	0	0
HCCS (2)	0	-0.14 (0.39)	0	-0.08 (0.78)	0	-0.07 (0.62)	0	0
BLVRA (1)	0	-0.15 (0.19)	0	-0.16 (0.27)	0	-0.13 (0.29)	0	-0.18 (0.14)
CP	0.52 (0.03)	0	0.57 (0.01)	0	0.57 (0.01)	0	0.69 (0.003)	0
UROD (1)	0	0	0	0	0	0	0	0
UROD (2)	0	0	0	0	0	0	0	0
BLVRA (2)	-0.03 (0.68)	0	0	0	0 (0.82)	0	0	0
ADH6	-0.24 (0.41)	0.53 (0.034)	-0.22 (0.63)	0.58 (0.04)	0	0.78 (0.001)	0	0.76 (0.001)
HMOX2	0	0.25 (0.22)	0	0.29 (0.34)	0	0.24 (0.26)	0	0.43 (0.04)

8. Discussion

We proposed two BIC criteria for selecting the tuning parameter. The first resembles the one used in univariate variable selection, and is used in estimating $\boldsymbol{\beta}$ with a fixed, or consistently estimated precision matrix. It is consistent for model selection in both cases when an oracle penalty function is used. Our proof allows both p and n go to infinity. The second BIC criterion is used when $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are jointly estimated, as in the two-stage procedure. Simulation results show that the second BIC performs well, and outperforms the tuning parameter selection procedure using an external data set. Future research is needed for theoretical results for this BIC, and for $p > n$.

Our primary focus in this paper is on variable selection for regression coefficients. The optimal regression coefficient estimator that is consistent for model selection and has the oracle properties only requires a consistent estimator of the precision matrix, e.g., using GLASSO. Future research is needed to develop a more efficient estimator of the precision matrix. In the non-regression setting assuming y has mean 0, $\log(m)/n \to 0$, and sub-gaussian errors, Lam and Fan (2009) provided the rate of convergence for the $\boldsymbol{\Omega}$ estimator for Lasso and oracle penalized estimators. Their results could be extended to regression settings in the presence of covariates X for $p_0/n \to 0$. When $p_0 \gg n$, say $\log(p_0)/n \to 0$, additional assumptions are needed on the design matrix and the sparsity of $\boldsymbol{\beta}$. This is a topic for future research.

Acknowledgement

This work is supported by the NCI grants R37 CA076404 and P01 CA134294.

References

Arias-Castro, E., Candès, E. J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39**, 2533-2556.

Bhatia, R. (1997). *Matrix Analysis*. Springer-Verlag. New York.

Brown, P. J., Frean, R. and Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika* **86**, 635-648.

Dicker, L., Huang, B. and Lin, X. (2013). Variable selection and estimation with the seamless-L_0 penalty. *Statist. Sinica* **23**, 929-962.

Fan, J. and Li, R. (2001). Variable selection via nonconvace penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

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

Gao, X., Pu, Q., Wu, Y. and Xu, H. (2009). Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. Arxiv preprint arXiv:0909.0934.

Lam, C., and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254-4278.

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *J. Statist. Plann. Inference* **141**, 2839-2848.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

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

Mootha, V. K. et al. (2003). PGC-1α-responsive Genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273.

Peng, J. et al., (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Statist.* **4**, 53-77.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic J. Statist.* **2**, 595-515.

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional Genomics. *Statist. Appl. Genet. Mol. Biol.* **4**, Article 32.

Stewart, G. W. and Sun, J. (1990). *Matrix perturbation theory.* Academic Press, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

Wang, H, Li, B, and Leng, C. (2009). Shrinkage tuning parameters selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 617-683.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **95**, 19-35.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Ann. Statist.* **38**, 894-942.

Department of Biostatistics, University of Washington, Seattle, 98105, USA.

E-mail: tsofer@uw.edu

Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: ldicker@stat.rutgers.edu

Department of Biostatistics, Harvard School of Public Health, Boston, MA, 02115, USA.

E-mail: xlin@hsph.harvard.edu

(Received January 2013; accepted December 2013)