Shrinking characteristics of precision matrix estimators

Aaron J. Molstad* and Adam J. Rothman
School of Statistics, University of Minnesota

May 10, 2017

Abstract

We propose a framework to shrink a user-specified characteristic of a precision matrix estimator that is needed to fit a predictive model. Estimators in our framework minimize the Gaussian negative log-likelihood plus an L_1 penalty on a linear or affine function evaluated at the optimization variable corresponding to the precision matrix. We establish convergence rate bounds for these estimators and we propose an alternating direction method of multipliers algorithm for their computation. Our simulation studies show that our estimators can perform better than competitors when they are used to fit predictive models. In particular, we illustrate cases where our precision matrix estimators perform worse at estimating the population precision matrix while performing better at prediction.

1 Introduction

Estimating precision matrices is required to fit many statistical models. Many papers written in the last decade have proposed shrinkage estimators of the precision matrix when p, the number of variables, is large. Pourahmadi (2013) and Fan et al. (2016) provide comprehensive reviews of large covariance and precision matrix estimation. The main strategy used in many of these papers is minimize the Gaussian negative log-likelihood plus a penalty on the off-diagonal entries of the optimization variable corresponding to the precision matrix. For example, Yuan and Lin (2007) proposed the L_1-penalized Gaussian likelihood precision matrix estimator defined by

$$\underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \mathrm{tr}(S\Omega) - \log\det(\Omega) + \lambda \sum_{i \neq j} |\Omega_{ij}| \right\}, \tag{1}$$

where S is the sample covariance matrix, $\lambda > 0$ is a tuning parameter, $\mathbb{S}_+^p$ is the set of $p \times p$ symmetric and positive definite matrices, and tr and $\det$ are the trace and determinant, respectively. Other authors have replaced the L_1 penalty in (1) with the squared Frobenius norm (Witten and Tibshirani, 2009; Rothman and Forzani, 2014) or non-convex penalties that also encourage zeros in the estimator (Lam and Fan, 2009; Fan et al., 2009).

To fit many predictive models, only a characteristic of the population precision matrix needs to be estimated. For example, in binary linear discriminant analysis, the population precision matrix is

*Corresponding author: molst029@umn.edu

needed for prediction only through the product of the precision matrix and the difference between the two conditional distribution mean vectors. Many authors have proposed methods that directly estimate this characteristic (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012).

We propose to estimate the precision matrix by shrinking the characteristic of the estimator that is needed for prediction. The characteristic we consider is a linear or affine function evaluated at the precision matrix. The goal is to improve prediction performance. Unlike methods that estimate the characteristic directly, our approach provides the practitioner an estimate of the entire precision matrix, not just the characteristic. In our simulation studies and data example, we show that penalizing the characteristic needed for prediction can improve prediction performance over competing sparse precision estimators like (1), even when the true precision matrix is very sparse. In addition, estimators in our framework can be used in applications other than linear discriminant analysis.

2 Proposed method

2.1 Penalized likelihood estimator

We propose to estimate the population precision matrix Ω_* with

$$\hat{\Omega} = \underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda |A\Omega B - C|_1 \right\}, \tag{2}$$

where $A \in \mathbb{R}^{a \times p}$, $B \in \mathbb{R}^{p \times b}$, and $C \in \mathbb{R}^{a \times b}$ are user-specified matrices; and $|M|_1 = \sum_{i,j} |M_{ij}|$. Our estimator exploits the assumption that $A\Omega_* B - C$ is sparse. In cases where A, B, and C need to be estimated, we replace them with their estimators.

An estimator defined by (2) with $C = 0$ was mentioned in an unpublished manuscript by Dalal and Rajaratnam (2014) available on arXiv. These authors proposed an alternating minimization algorithm for solving (1) and described how to apply it to solve (2). Dalal and Rajaratnam did not describe applications or theoretical properties of this estimator. Also, as written, their algorithm does not actually solve (2) when A and B are arbitrary. We propose an alternating direction method of multipliers algorithm to solve (2) and establish theoretical properties for this estimator.

2.2 Example applications

Fitting the discriminant analysis model requires the estimation of one or more precision matrices. In particular, the linear discriminant analysis model assumes that the data are independent copies of the random pair (X, Y), where the support of Y is $\{1, \ldots, J\}$ and

$$X \mid Y = j \sim \mathrm{N}_p\left(\mu_{*j}, \Omega_*^{-1}\right), \quad j = 1, \ldots, J, \tag{3}$$

where $\mu_{*j} \in \mathbb{R}^p$ and $\Omega_*^{-1} \in \mathbb{S}_+^p$ are unknown. To discriminate between response categories l and m, only the characteristic $\Omega_*(\mu_{*l} - \mu_{*m})$ is needed. Methods that estimate this characteristic directly have been proposed (Cai and Liu, 2011; Mai et al., 2012; Fan et al., 2012; Mai et al., 2015). These methods are useful in high dimensions because they perform variable selection. For the jth variable to be non-informative for discriminating between response categories l and m, it must be that the jth element of $\Omega_*(\mu_{*l} - \mu_{*m})$ is zero. While these methods can perform well in classification and variable selection, they do not actually fit the model in (3).

Methods for fitting (3) specifically for linear discriminant analysis either assume Ω_* is diagonal (Bickel and Levina, 2004) or that both $\mu_{*l} - \mu_{*m}$ and Ω_* are sparse (Guo, 2010; Xu et al., 2015). A method for fitting (3) and performing variable selection was proposed by Witten and Tibshirani (2009). They suggest a two-step procedure where one first estimates Ω_*, and then with the estimate $\bar{\Omega}$ fixed, estimates each μ_{*j} by penalizing the characteristic $\bar{\Omega}\mu_j$, where μ_j is the optimization variable corresponding to μ_{*j}.

To apply our method to the linear discriminant analysis problem, we use (2) with $A = I_p$, $C = 0$, and B equal to the matrix whose columns are $\bar{x}_j - \bar{x}_k$ for all $1 \leq j < k \leq J$, where $\bar{x}_j$ is the unpenalized maximum likelihood estimator of μ_{*j}. For large values of the tuning parameter, this would lead an estimator of Ω_* such that $\hat{\Omega}(\bar{x}_j - \bar{x}_k)$ is sparse. Thus our approach simultaneously fits (3) and performs variable selection.

Precision and covariance matrix estimators are also needed for portfolio allocation. The optimal allocation based on the Markowitz (1952) minimum-variance portfolio is proportional to $\Omega_*\mu_*$, where μ_* is the vector of expected returns for p assets and Ω_* is precision matrix for the returns. In practice, one would estimate Ω_* and μ_* with their usual sample estimators $\hat{\Omega}$ and $\hat{\mu}$. However, when p is large, the usual sample estimator of Ω_* does not exist, so regularization is necessary. Moreover, Brodie et al. (2009) argue that sparse portfolios are often desirable when p is large. While many have proposed using sparse or shrinkage estimators of Ω_* or Ω_*^{-1} plugged-in to the Markowitz criterion, e.g., Xue et al. (2012), this would not necessarily lead to sparse estimators of $\Omega_*\mu_*$. Chen et al. (2016) proposed a method for estimating the characteristic $\Omega_*\mu_*$ directly, but like the direct linear discriminant methods, this approach does not lead to an estimate of Ω_*. For the sparse portfolio allocation problem, we propose to estimate Ω_* using (2) with $A = I_p$, $C = 0$, and $B = \hat{\mu}$.

Another application is in linear regression where the response and predictor have a joint multivariate normal distribution. In this case, the regression coefficient matrix is $\Omega_*\Sigma_{*XY}$, where Ω_* is the marginal precision matrix for the predictors and Σ_{*XY} is the cross-covariance matrix between predictors and responses. We propose to estimate Ω_* using (2) with $A = I_p$, $C = 0$, and B equal to the usual sample estimator of Σ_{*XY}. Similar to the procedure proposed by Witten and Tibshirani (2009), this approach provides an alternative method for estimating regression coefficients using shrinkage estimators of the marginal precision matrix for the predictors.

3 Computation

3.1 Alternating direction method of multipliers algorithm

To solve the optimization in (2), we propose an alternating direction method of multipliers algorithm with a modification based on the majorize-minimize principle (Lange, 2016). Following the standard alternating direction method of multipliers approach (Boyd et al., 2011), we rewrite (2) as a constrained optimization problem:

$$\underset{(\Theta,\Omega)\in\mathbb{R}^{a\times b}\times\mathbb{S}_+^p}{\arg\min} \{\operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda|\Theta|_1\} \quad \text{subject to } A\Omega B - \Theta = C. \tag{4}$$

The augmented Lagrangian for (4) is defined by

$$\mathcal{F}_\rho(\Omega, \Theta, \Gamma) = \operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda|\Theta|_1 \\ - \operatorname{tr}\left\{\Gamma^{\mathrm{T}}(A\Omega B - \Theta - C)\right\} + \frac{\rho}{2}\|A\Omega B - \Theta - C\|_F^2,$$

where $\rho > 0$ and $\Gamma \in \mathbb{R}^{a\times b}$ is the Lagrangian dual variable. Let the subscript k denote the kth iterate. From Boyd et al. (2011), to solve (4), the alternating direction method of multipliers algorithm uses the following updating equations:

$$\Omega_{k+1} = \underset{\Omega\in\mathbb{S}_+^p}{\arg\min}\, \mathcal{F}_\rho(\Omega, \Theta_k, \Gamma_k), \tag{5}$$

$$\Theta_{k+1} = \underset{\Theta\in\mathbb{R}^{a\times b}}{\arg\min}\, \mathcal{F}_\rho(\Omega_{k+1}, \Theta, \Gamma_k), \tag{6}$$

$$\Gamma_{k+1} = \Gamma_k - \rho\left(A\Omega_{k+1}B - \Theta_{k+1} - C\right). \tag{7}$$

Instead of solving (5) exactly, we approximate its objective function with a majorizing function. Specifically, we replace (5) with

$$\Omega_{k+1} = \underset{\Omega\in\mathbb{S}_p^+}{\arg\min}\left\{\mathcal{F}_\rho(\Omega, \Theta_k, \Gamma_k) + \frac{\rho}{2}\mathrm{vec}(\Omega - \Omega_k)^{\mathrm{T}} Q \mathrm{vec}(\Omega - \Omega_k)\right\}, \tag{8}$$

where $Q = \tau I - \left(A^{\mathrm{T}}A \otimes BB^{\mathrm{T}}\right)$, τ is selected so that $Q \in \mathbb{S}_+^p$, $\otimes$ is the Kronecker product, and vec forms a vector by stacking the columns of its matrix argument. Since

$$\mathrm{vec}(\Omega - \Omega_k)^{\mathrm{T}}\left(A^{\mathrm{T}}A \otimes BB^{\mathrm{T}}\right)\mathrm{vec}(\Omega - \Omega_k) = \operatorname{tr}\left\{A^{\mathrm{T}}A(\Omega - \Omega_k)BB^{\mathrm{T}}(\Omega - \Omega_k)\right\},$$

we can rewrite (8) as

$$\Omega_{k+1} = \underset{\Omega\in\mathbb{S}_p^+}{\arg\min}\left[\mathcal{F}_\rho(\Omega, \Theta_k, \Gamma_k) + \frac{\rho\tau}{2}\|\Omega - \Omega_k\|_F^2 - \frac{\rho}{2}\operatorname{tr}\left\{A^{\mathrm{T}}A(\Omega - \Omega_k)BB^{\mathrm{T}}(\Omega - \Omega_k)\right\}\right],$$

which is equivalent to

$$\Omega_{k+1} = \underset{\Omega\in\mathbb{S}_p^+}{\arg\min}\left[\operatorname{tr}\left\{\left(S + G_k\right)\Omega\right\} - \log\det(\Omega) + \frac{\rho\tau}{2}\|\Omega - \Omega_k\|_F^2\right], \tag{9}$$

where $G_k = \rho A^{\mathrm{T}}(A\Omega_k B - \rho^{-1}\Gamma_k - \Theta_k - C)B^{\mathrm{T}}$. The zero gradient equation for (9) is

$$S - \Omega_{k+1}^{-1} + \frac{1}{2}\left(G_k + G_k^{\mathrm{T}}\right) + \rho\tau\left(\Omega_{k+1} - \Omega_k\right) = 0, \tag{10}$$

which is a quadratic equation that can be solved in closed form (Witten and Tibshirani, 2009; Price et al., 2015). The solution is

$$\Omega_{k+1} = \frac{1}{2\rho\tau}U\left\{-\Psi + \left(\Psi^2 + 4\rho\tau I_p\right)^{1/2}\right\}U^{\mathrm{T}},$$

where $U\Psi U^{\mathrm{T}}$ is the eigendecomposition of $S + 2^{-1}(G_k + G_k^{\mathrm{T}}) - \rho\tau\Omega_k$. Our majorize-minimize approach is a special case of the *prox-linear* alternating direction method of multiplier algorithm (Chen and Teboulle, 1994; Deng and Yin, 2016).

Conveniently, (6) also has a closed form solution:

$$\Theta_{k+1} = \text{soft}\left(A\Omega_{k+1}B - \rho^{-1}\Gamma_k - C, \rho^{-1}\lambda\right),$$

where $\text{soft}(x, \tau) = \max\left(|x| - \tau, 0\right)\text{sign}(x)$. To summarize, we solve (2) with the following algorithm.

Algorithm 1 Alternating direction method of multipliers algorithm for (2)

Initialize $\Omega_{(0)} \in \mathbb{S}_+^p$, $\Theta_{(0)} \in \mathbb{R}^{a\times b}$, $\rho > 0$, and τ such that Q is positive definite. Set $k = 0$. Repeat Step 1 - 6 until convergence:

Step 1. Compute $G_k = \rho A^{\mathrm{T}}(A\Omega_k B - \rho^{-1}\Gamma_k - \Theta_k - C)B^{\mathrm{T}}$;
Step 2. Decompose $S + 2^{-1}(G_k + G_k^{\mathrm{T}}) - \rho\tau\Omega_k = U\Psi U^{\mathrm{T}}$ where U is orthogonal and Ψ is diagonal;
Step 3. Set $\Omega_{k+1} = (2\rho\tau)^{-1}U\left\{-\Psi + (\Psi^2 + 4\rho\tau I_p)^{1/2}\right\}U^{\mathrm{T}}$;
Step 4. Set $\Theta_{k+1} = \text{soft}(A\Omega_{k+1}B - \rho^{-1}\Gamma_k - C, \rho^{-1}\lambda)$;
Step 5. Set $\Gamma_{k+1} = \Gamma_k - \rho\left(A\Omega_{k+1}B - \Theta_{k+1} - C\right)$;
Step 6. Replace k with $k + 1$.

3.2 Convergence and implementation

Using the same proof technique as in Deng and Yin (2016), one can show that the iterates from Algorithm 1 converge to their optimal values when a solution to (4) exists.

In our implementation, we set $\tau = \varphi_1(A^{\mathrm{T}}A)\varphi_1(BB^{\mathrm{T}}) + 10^{-8}$, where $\varphi_1(\cdot)$ denotes the largest eigenvalue of its argument. This computation is only needed once at the initialization of our algorithm. We expect that in practice, the computational complexity of our algorithm will be dominated by the eigendecomposition in Step 2, which requires $O(p^3)$ flops.

To select the tuning parameter to use in practice, we recommend using some type of cross-validation procedure based on the application. For example, in the linear discriminant analysis case, one could select the tuning parameter that minimizes the validation misclassification rate or maximizes a validation likelihood.

4 Statistical Properties

In this section, we show that by using the penalty in (2), we can estimate Ω_* and $A\Omega_*B$ consistently in the Frobenius and L_1 norms, respectively. Our results rely on assuming that $A\Omega_*B$ is sparse. Define the set $\mathcal{G}$ as the indices of $A\Omega_*B$ that are nonzero, i.e.,

$$\mathcal{G} = \left\{(i,j) \in \{1,\ldots,a\} \times \{1,\ldots,b\} : [A\Omega_*B]_{ij} \neq 0\right\}.$$

Let the notation $[A\Omega_*B]_{\mathcal{G}} \in \mathbb{R}^{a\times b}$ denote the matrix whose (i,j)th entry is equal to the (i,j)th of $A\Omega_*B$ if $(i,j) \in \mathcal{G}$ and is equal to zero if $(i,j) \notin \mathcal{G}$. We generalize our results to the case that A

and B are unknown, and we use plug-in estimators of them in (2).

We first establish convergence rates for the case that A and B are known. Let $\sigma_j(\cdot)$ and $\varphi_j(\cdot)$ denote the jth largest singular value and eigenvalue of their arguments respectively. Suppose that the sample covariance matrix used in (2) is $S_n = n^{-1}\sum_{i=1}^n X_i X_i^{\mathrm{T}}$, where $X_1, \ldots, X_n$ are independent and identically distributed p_n-dimensional random vectors with mean zero and covariance matrix Ω_*^{-1}. We will make the following assumptions:

Assumption 1. *For all n, there exists a constant k_1 such that*

$$0 < k_1^{-1} \leq \varphi_{p_n}(\Omega_*) \leq \varphi_1(\Omega_*) \leq k_1 < \infty.$$

Assumption 2. *For all n, there exists a constant k_2 such that* $\min\{\sigma_{p_n}(A), \sigma_{p_n}(B)\} \geq k_2 > 0$.

Assumption 3. *For all n, there exist positive constants k_3 and k_4 such that*

$$\max_{j\in\{1,\ldots,p_n\}} E\left\{\exp(tX_{1j}^2)\right\} \leq k_3 < \infty \textit{ for all } t \in (-k_4, k_4).$$

Assumptions 1 and 3 are common in the regularized precision matrix estimation literature, e.g., Assumption 1 was made by Bickel and Levina (2008), Rothman et al. (2008) and Lam and Fan (2009) and Assumption 3 holds if X_1 is multivariate Normal. Assumption 2 requires that A and B are both rank p_n, which has the effect of shrinking every entry of $\hat{\Omega}$. The convergence rate bounds we establish also depend on the quantity

$$\xi(p_n, \mathcal{G}) = \sup_{M\in\mathbb{S}^{p_n}, M\neq 0} \frac{|\,[AMB]_{\mathcal{G}}\,|_1}{\|M\|_F},$$

where $\mathbb{S}^{p_n}$ is the set of symmetric $p_n \times p_n$ matrices. Negahban et al. (2012) defined a similar and more general quantity and called it a compatibility constant.

Theorem 1. *Suppose Assumptions 1–3 are true. If $\lambda_n = K_1(n^{-1}\log p_n)^{1/2}$, K_1 is sufficiently large, and $\xi^2(p_n, \mathcal{G})\log p_n = o(n)$, then (i) $\|\hat{\Omega} - \Omega_*\|_F = O_P\{\xi(p_n, \mathcal{G})(\log p_n/n)^{1/2}\}$ and (ii) $|A\hat{\Omega}B - A\Omega_* B|_1 = O_P\{\xi^2(p_n, \mathcal{G})(\log p_n/n)^{1/2}\}$.*

The quantity $\xi(p_n, \mathcal{G})$ can be used to recover known results for special cases of (2). For example, when A and B are identity matrices, $\xi(p_n, \mathcal{G}) = {s_n}^{1/2}$, where s_n is the number of nonzero entries in Ω_*. This special case was established by Rothman et al. (2008). We can simplify the results of Theorem 1 for case that $A\Omega_* B$ has g_n nonzero entries by introducing an additional assumption:

Assumption 4. *For all n, there exists a constant k_5 such that*

$$\sup_{M\in\mathbb{S}^{p_n}, M\neq 0} \frac{\|[AMB]_{\mathcal{G}}\|_F}{\|M\|_F} \leq k_5 < \infty.$$

Assumption 4 is not the same as bounding $\xi(p_n, \mathcal{G})$ because the numerator uses the Frobenius norm instead of the L_1 norm. This requires that for those entries of $A\Omega_* B$ which are nonzero, the corresponding rows and columns of A and B, respectively, do not have magnitudes too large as p_n grows.

Remark 1. *Assume that the conditions of Theorem 1 are true. If Assumption 4 is true and $A\Omega_* B$ has g_n nonzero entries, then $\|\hat{\Omega} - \Omega_*\|_F = O_P\{(g_n \log p_n/n)^{1/2}\}$ and $|A\hat{\Omega}B - A\Omega_* B|_1 = O_P\{(g_n^2 \log p_n/n)^{1/2}\}$.*

In practice, A and B are often unknown and must be estimated. Let $\hat{A}_n$ and $\hat{B}_n$ be estimators of A and B. In this case, we estimate $A\Omega_* B$ with $\hat{A}_n\tilde{\Omega}\hat{B}_n$, where

$$\tilde{\Omega} = \underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \mathrm{tr}(S_n\Omega) - \log\det(\Omega) + \lambda_n |\hat{A}_n \Omega \hat{B}_n|_1 \right\}. \tag{11}$$

Suppose that there exist sequences a_n and b_n such that $|(A - \hat{A}_n)A^+|_1 = O_P(a_n)$ and $|B^+(B - \hat{B}_n)|_1 = O_P(b_n)$, where A^+ and B^+ are the Moore–Penrose pseudoinverses of A and B, respectively; and $a_n = o(1)$, and $b_n = o(1)$. Let $C_n = \max\{a_n, b_n\} \,|\, [A\Omega_* B]_{\mathcal{G}}\,|_1$.

Theorem 2. *Suppose Assumptions 1–3 are true. If $\lambda_n = K_2(n^{-1}\log p_n)^{1/2}$, K_2 is sufficiently large, $\max\{a_n, b_n\} = o(1)$, $\xi^2(p_n, \mathcal{G})\log p_n = o(n)$, and $C_n^2 \log p_n = o(n)$, then*
(i) $\|\tilde{\Omega} - \Omega\|_F = O_P\{\xi(p_n, \mathcal{G})(\log p_n/n)^{1/2} + C_n^{1/2}(\log p_n/n)^{1/4}\}$ and
(ii) $|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 = O_P\{\xi^2(p_n, \mathcal{G})(\log p_n/n)^{1/2} + C_n^{1/2}\xi(p_n, \mathcal{G})(\log p_n/n)^{1/4} + C_n\}$.

The convergence rate bounds in Theorem 2 are the sum of the statistical errors from Theorem 1 plus additional errors which comes from estimating A and B.

5 Simulation studies

5.1 Models

We compare our precision matrix estimator to competing estimators when they are used to fit the linear discriminant analysis model. For 100 independent replications, we generated a realization of n independent copies of (X, Y) defined in (3), where $\mu_{*j} = \Omega_*^{-1}\beta_{*j}$ and $P(Y = j) = 1/J$ for $j = 1, \ldots, J$. Using this construction, if the kth element of $\beta_{*l} - \beta_{*m}$ is zero, i.e., $\Omega_*(\mu_{*l} - \mu_{*m})$ is zero, then the kth variable is non-informative for discriminating between response categories l and m.

For each $J \in \{3, \ldots, 10\}$, we partition our n observations into a training set of size $25J$, a validation set of size 200, and a test set of size 1000. We considered two models for Ω_*^{-1} and β_{*j}. Let $\mathbb{1}(\cdot)$ be the indicator function.

Model 1. We set $\beta_{*j,k} = 1.5\,\mathbb{1}\left[k \in \{4(j-1)+1, \ldots, 4j\}\right]$, so that for any pair of response categories, only eight variables were informative for discrimination. We set $\Omega_{*a,b}^{-1} = .9^{|a-b|}$, so that Ω_* was tridiagonal.

Model 2. We set $\beta_{*j,k} = 2\,\mathbb{1}\left[k \in \{5(j-1)+1, \ldots, 5j\}\right]$, so that for any pair of response categories, only ten variables were informative for discrimination. We set Ω_*^{-1} to be block diagonal: the block corresponding to the informative variables, i.e., the first $5J$ variables, had off diagonal entries equal to 0.5 and diagonal entries equal to one. The block submatrix corresponding to the $p - 5J$ non-informative variables had (a, b)th entry equal to $0.5^{|a-b|}$.

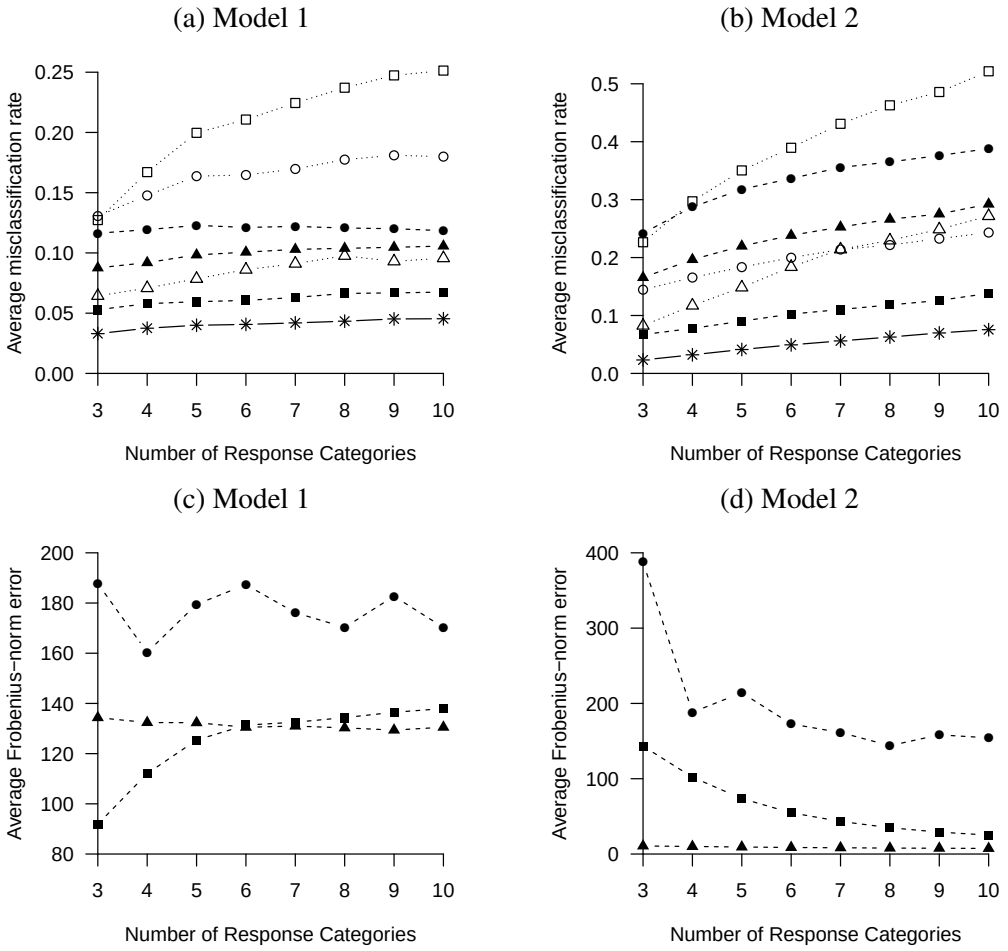

Figure 1: Misclassification rates and Frobenius norm error averaged over 100 replications with $p = 200$ for Models 1 and 2. The methods displayed are the estimator we proposed in Section 2.2 (dashed and ■), the L_1-penalized Gaussian likelihood estimator (dashed and ▲), the Ledoit-Wolf-type estimator from (12) (dashed and ●), Bayes (solid and $*$), the method proposed by Guo (2010) (dots and ○), the method proposed by Mai et al. (2015) (dots and △), and the method proposed by Witten and Tibshirani (2011) (dots and □).

For both models, sparse estimators of Ω_* should perform well because the precision matrices are very sparse. The total number of informative variables is $4J$ and $5J$ in Models 1 and 2 respectively, so a method like that proposed by Mai et al. (2015), which selects variables that are informative for all pairwise comparisons, may perform poorly when J is large.

5.2 Methods

We compared several methods in terms of classification accuracy on the test set. We fit (3) using the following methods: the sparse naïve Bayes estimator proposed by Guo (2010) with tuning parameter chosen to minimize misclassification rate on the validation set; and the Bayes rule, i.e., Ω_*, μ_{*j}, and $P(Y = j)$ known for $j = 1, \ldots, J$. We also fit (3) using the ordinary sample means and using the following precision matrix estimators: the estimator we proposed in Section

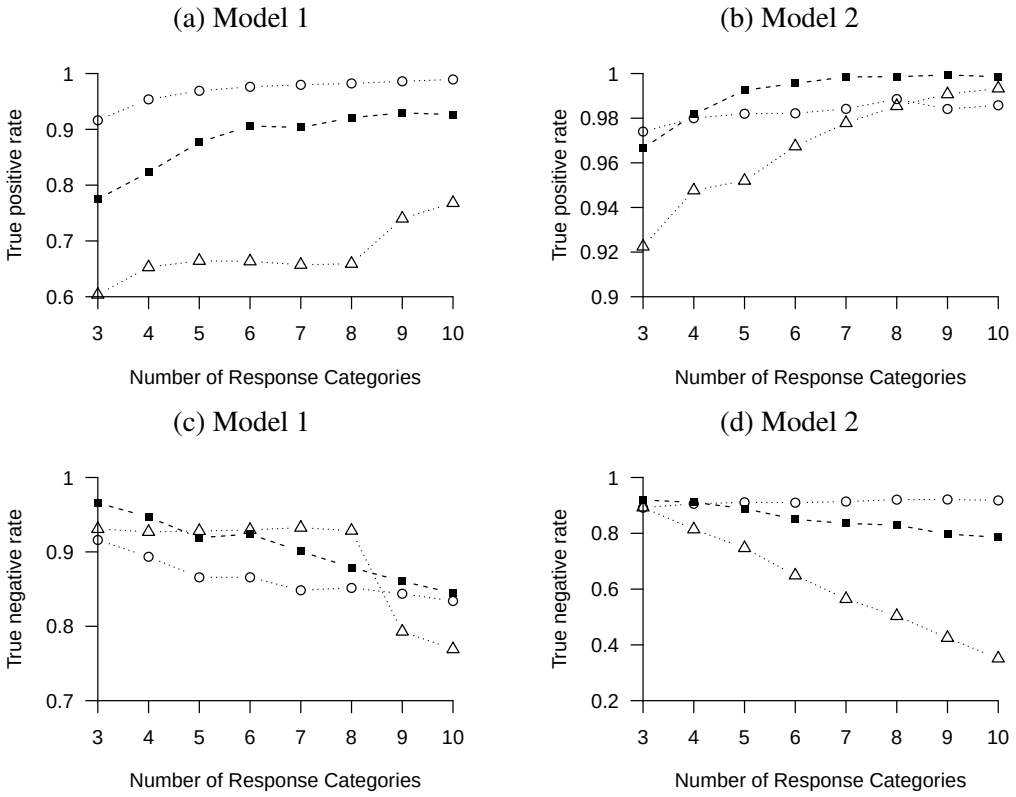

Figure 2: True positive and true negative rates averaged over 100 replications with $p = 200$ for Model 1 in (a) and (c); and for Model 2 in (b) and (d). The methods displayed are the estimator we proposed in Section 2.2 (dashed and ■), the method proposed by Guo (2010) (dots and ○), and the method proposed by Mai et al. (2015) (dots and △).

2.2 with tuning parameter chosen to minimize misclassification rate on the validation set; the L_1-penalized Gaussian likelihood precision matrix estimator (Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008) with the tuning parameter chosen to minimize the misclassification rate of the validation set; and a covariance matrix estimator similar to the estimator proposed by Ledoit and Wolf (2004), which is defined by

$$\hat{\Omega}_{\mathrm{LW}}^{-1} = \alpha S + \gamma(1-\alpha) I_p, \tag{12}$$

where $(\alpha, \gamma) \in (0,1) \times (0,\infty)$ were chosen to minimize the misclassification rate of the validation set. The L_1-penalized Gaussian likelihood precision matrix estimator we used penalized the diagonals. With our data generating models, we found this performed better at classification than (1), which does not penalize the diagonals. We also tried two Fisher-linear-discriminant-based methods applicable to multi-category linear discriminant analysis: the sparse linear discriminant method proposed by Witten and Tibshirani (2011) with tuning parameter and dimension chosen to minimize the misclassification rate of the validation set; and the multi-category sparse linear discriminant method proposed by Mai et al. (2015) with tuning parameter chosen to minimize the misclassification rate of the validation set.

We could also have selected tuning parameters for the model-based methods by maximiz-

ing a validation likelihood or using an information criterion, but minimizing the misclassification rate on a validation set made it fairer to compare the model-based methods and the Fisher-linear-discriminant-based methods in terms of classification accuracy.

5.3 Performance measures

We measured classification accuracy on the test set for each replication for the methods described in Section 5.2. For the methods that produced a precision matrix estimator, we also measured this estimator's Frobenius norm error: $\|\bar{\Omega} - \Omega_*\|_F$, where $\bar{\Omega}$ is the estimator. To measure variable selection accuracy, we used both the true positive rate and the true negative rate, which are respectively defined by

$$\frac{\text{card}\left\{(m,k) : \hat{\Delta}_{m,k} \neq 0 \cap \Delta_{*m,k} \neq 0\right\}}{\text{card}\left\{(m,k) : \Delta_{*m,k} \neq 0\right\}}, \quad \frac{\text{card}\left\{(m,k) : \hat{\Delta}_{m,k} = 0 \cap \Delta_{*m,k} = 0\right\}}{\text{card}\left\{(m,k) : \Delta_{*m,k} = 0\right\}},$$

where $(m,k) \in \{2,\ldots,J\} \times \{1,\ldots,p\}$, $\Delta_{*m} = \beta_{*1} - \beta_{*m}$, $\hat{\Delta}_m$ is an estimator of Δ_{*m}, and card denotes the cardinality of a set.

5.4 Results

We display average misclassification rates and Frobenius norm error averages for both models with $p = 200$ in Figure 1, and display variable selection accuracy averages in Figure 2. For both models, our method outperformed all competitors in terms of classification accuracy for all J, except the Bayes rule, which uses population parameter values unknown in practice. In terms of precision matrix estimation, for Model 1, our method did better than the L_1-penalized Gaussian likelihood precision matrix estimator when the sample size was small, but became worse than the L_1-penalized Gaussian likelihood precision matrix estimator as the sample size increased. For Model 2, our method was worse than the L_1-penalized Gaussian likelihood precision matrix estimator in Frobenius norm error for precision matrix estimation, but was better in terms of classification accuracy.

In terms of variable selection, our method was competitive with the methods proposed by Guo (2010) and Mai et al. (2015). For Model 1, our method tended to have a higher average true negative rate than the method of Guo (2010) and a lower average true positive rate than the method of Mai et al. (2015). For Model 2, all methods tended to have relatively high average true positive rates, while our method had a higher average true negative rate than the method of Mai et al. (2015). Although the method proposed by Guo (2010) had a higher average true negative rate for Model 2 than our proposed method had, our method performed better in terms of classification accuracy.

6 Genomic data example

We used our method to fit the linear discriminant analysis model in a real data application. The data are gene expression profiles consisting of $p = 22,283$ genes from 127 subjects, who either have Crohn's disease, ulcerative colitis, or neither. This dataset comes from Burczynski et al. (2006). The goal of our analysis was to fit a linear discriminant analysis model that could be used to identify which genes are informative for discriminating between each pair of the response categories. These data were also analyzed in Mai et al. (2015). One difference between our method and the method

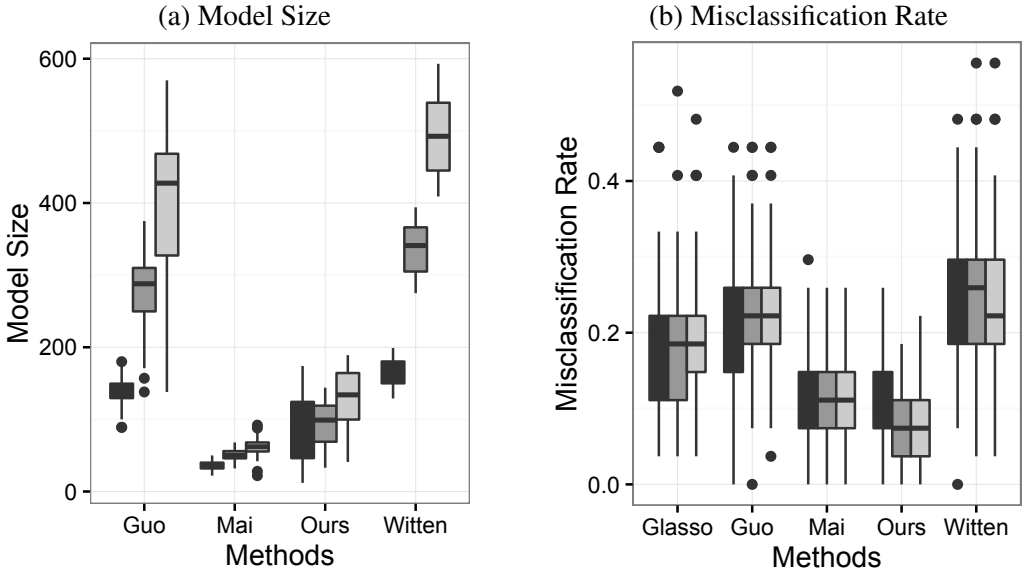

Figure 3: Model sizes and misclassification rates from 100 random training/testing splits with $k = 100$ (dark grey), $k = 200$ (grey), and $k = 300$ (light grey). Guo is the method proposed by Guo (2010), Mai is the method proposed by Mai et al. (2015), Glasso is the L_1-penalized Gaussian likelihood precision matrix estimator, Ours is the estimator we propose Section 2.2, and Witten is the method proposed by Witten and Tibshirani (2011).

of Mai et al. (2015) is that the method of Mai et al. (2015) excludes variables from all pairwise response category comparisons whereas our method allows a distinct set of informative variables to be estimated for each comparison.

To measure the classification accuracy of our method and its competitors, we randomly split the data into training set of size 100 and test set of size 27 for 100 independent replications. Within each replication, we first applied a screening rule to the training set as in Rothman et al. (2009) and Mai et al. (2015) based on F-test statistics, and then restricted our discriminant analysis model to the genes with the k largest F-test statistic values.

We chose tuning parameters with 5-fold cross validation that minimized the validation classification error rate. Misclassification rates are shown in Figure 3, where we compared our method to the method proposed by Mai et al. (2015), the method proposed by Witten and Tibshirani (2011), the method proposed by Guo (2010), and the method that used the L_1-penalized Gaussian likelihood precision matrix estimator. We saw that our method was as or more accurate in terms of classification accuracy than the competing methods. The only method that performed nearly as well was that of Mai et al. (2015) when we used $k = 100$ screened genes. However, the best out-of-sample classification accuracy was achieved with $k = 300$, where our method was significantly better than the competitors.

In Figure 3, we also display model sizes, i.e., the total number of variables that were estimated to be informative for discriminating between response categories. To measure model size for the method proposed by Witten and Tibshirani (2011), we used the version of their method with two discriminant vectors. We saw that although the method of Mai et al. (2015) tended to estimate slightly smaller models, our method, which performs best in classification, selects only slightly

more variables. Moreover, our method can be used to identify a distinct subset of genes that are informative specifically for discriminating between patients with Crohn's disease and ulcerative colitis. This was of interest in the study of Burczynski et al. (2006).

7 Acknowledgements

A. J. Molstad's research is supported by the Doctoral Dissertation Fellowship from the University of Minnesota. A. J. Rothman's research is supported in part by the National Science Foundation.

Appendix

Notation

Define the following norms: $\|A\|_\infty = \max_{i,j} |A_{ij}|$, $|A|_1 = \sum_{i,j} |A_{ij}|$, $\|A\|_F = \text{tr}(A^{\text{T}}A)$, $\|A\| = \sigma_1(A)$. Let $\mathbb{S}^p$ denote the set of $p \times p$ symmetric matrices. To simplify notation, we omit the subscript n from S_n, λ_n, and p_n as defined in 4 and let $\kappa = k_1^{-2}$.

Proof of Theorem 1

To prove Theorem 1, we use a strategy similar to that employed by Rothman (2012).

Lemma 1. *Suppose that Assumptions 1–3 hold, and* $\lambda \leq \epsilon\kappa \left\{\xi(p, \mathcal{G})\tau\right\}^{-1}$ *for some* $\tau > 12$. *Then for all positive and sufficiently small* ϵ, $\|B^+(S - \Omega_*^{-1})A^+\|_\infty \leq \lambda/2$ *implies* $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$.

Proof of Lemma 1: We follow the proof techniques used by Rothman et al. (2008), Negahban et al. (2012) and Rothman (2012). Define $B_\epsilon = \{\Delta \in \mathbb{S}^p : \|\Delta\|_F \leq \epsilon\}$. Let f be the objective function in (2). Because f is convex and $\hat{\Omega}$ is its minimizer, $\inf\{f(\Omega_* + \Delta) : \Delta \in B_\epsilon\} > f(\Omega_*)$, implies $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$ (Rothman et al., 2008). Define $D(\Delta) = f(\Omega_* + \Delta) - f(\Omega_*)$. Then

$$D(\Delta) = \text{tr}(S\Delta) + \log\det(\Omega_*) - \log\det(\Omega_* + \Delta) + \lambda_1 \left\{|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1\right\}.$$

By the arguments used in Rothman et al. (2008), $\log\det(\Omega_*) - \log\det(\Omega_* + \Delta) \geq -\text{tr}(S\Omega_*^{-1}) + 8^{-1}\kappa\|\Delta\|_F^2$, so that

$$D(\Delta) \geq \text{tr}\left\{\Delta(S - \Omega_*^{-1})\right\} + \frac{1}{8}\kappa\|\Delta\|_F^2 + \lambda_1 \left\{|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1\right\}. \tag{13}$$

We now bound $|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1$ in (13). Recall that

$$\mathcal{G} = \{(i,j) \in \{1, \ldots, a\} \times \{1, \ldots, b\} : [A\Omega_* B]_{ij} \neq 0\}$$

and $\mathcal{G}^c = \{1, \ldots, a\} \times \{1, \ldots, b\} \setminus \mathcal{G}$. Since $|A\Omega_* B|_1 = |\,[A\Omega_* B]_{\mathcal{G}}\,|_1$ and $|A(\Omega_* + \Delta)B|_1 = |\,[A\Omega_* B]_{\mathcal{G}} + [A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1$, we can apply the reverse triangle inequality: $|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1 \geq |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1$. Plugging this bound into (13),

$$D(\Delta) \geq \text{tr}\left\{(S - \Omega_*^{-1})\Delta\right\} + \frac{1}{8}\kappa\|\Delta\|_F^2 + \lambda_1 \left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right). \tag{14}$$

We now bound $\mathrm{tr}\left\{(S-\Omega_*^{-1})\Delta\right\}$. Let $A^+ = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ and $B^+ = B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}$. Because A and B are both rank p by Assumption 2, $A^+A = I_p$ and $BB^+ = I_p$. Thus

$$\begin{aligned}\mathrm{tr}\left\{(S-\Omega_*^{-1})\Delta\right\} &\geq -|\mathrm{tr}\left\{(S-\Omega_*^{-1})\Delta\right\}| = -|\mathrm{tr}\left\{(S-\Omega_*^{-1})A^+A\Delta BB^+\right\}| \\ &= -|\mathrm{tr}\left\{B^+(S-\Omega_*^{-1})A^+A\Delta B\right\}| \\ &\geq -\|B^+(S-\Omega_*^{-1})A^+\|_\infty |A\Delta B|_1. \end{aligned} \tag{15}$$

By assumption, $\|B^+(S-\Omega_*^{-1})A^+\|_\infty \leq \lambda/2$, so applying (15) to (14),

$$\begin{aligned} D(\Delta) &\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}|A\Delta B|_1 + \lambda\left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right) \qquad (16)\\ &= \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}\left(|\,[A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1\right) + \lambda\left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right) \\ &= \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 \\ &\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1. \qquad (17)\end{aligned}$$

We now bound the quantity $|\,[A\Delta B]_{\mathcal{G}}\,|_1$. Multiplying and dividing Δ by $\|\Delta\|_F$,

$$\left|\left[A\frac{|\Delta\|_F}{\|\Delta\|_F}\Delta B\right]_{\mathcal{G}}\right|_1 = \|\Delta\|_F\left|\left[A\frac{1}{\|\Delta\|_F}\Delta B\right]_{\mathcal{G}}\right|_1 \leq \|\Delta\|_F\left(\sup_{M\in\mathbb{S}^p, M\neq 0}\frac{|\,[AMB]_{\mathcal{G}}\,|_1}{\|M\|_F}\right),$$

so that $|\,[A\Delta B]_{\mathcal{G}}\,|_1 \leq \|\Delta\|_F\,\xi(p,\mathcal{G})$. Finally, since $\lambda \leq \epsilon\kappa\left\{\tau\xi(p,\mathcal{G})\right\}^{-1}$ with $\tau > 12$, $\|\Delta\|_F = \epsilon$ for $\Delta \in \mathcal{B}_\epsilon$,

$$\begin{aligned} D(\Delta) &\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda\|\Delta\|_F\xi(p,\mathcal{G}) \\ &= \|\Delta\|_F^2\left\{\frac{1}{8}\kappa - \frac{3\lambda\xi(p,\mathcal{G})}{2\|\Delta\|_F}\right\} \geq \epsilon^2\left(\frac{1}{12}\kappa - \frac{1}{\tau}\kappa\right) > 0. \end{aligned}$$

which establishes the desired result. □

The following lemma follows from the proof of Lemma 1 of Negahban et al. (2012).

Lemma 2. *If the conditions of Lemma 1 are true, then $\hat{\Delta} = \hat{\Omega} - \Omega_*$ belongs to the set*

$$\left\{\Delta \in \mathbb{S}^p : |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 \leq 3|\,[A\Delta B]_{\mathcal{G}}\,|_1\right\}.$$

Lemma 3 follows from the proof of Lemma 2 from Lam and Fan (2009), Assumption 2, and Lemma A.3 of Bickel and Levina (2008).

Lemma 3. *Suppose Assumptions 1–3 hold. Then, there exist constants C_1 and C_2 such that*

$$P(\|B^+SA^+ - B^+\Omega_*^{-1}A^+\|_\infty \geq \nu) \leq C_1p^2\exp(-C_2n\nu^2),$$

for $|\nu| \leq \delta$ where C_1, C_2, and δ do not depend on n.

Proof of Theorem 1: Set $\epsilon = K_1\kappa^{-1}\xi(p,\mathcal{G})(n^{-1}\log p)^{1/2}$, $\lambda = K_1\tau_1^{-1}(n^{-1}\log p)^{1/2}$ with $\tau_1 >$

12. Applying Lemma 1 and Lemma 3, there exist constants C_1 and C_2 such that for sufficiently large n,

$$P\left(\|\hat{\Omega}-\Omega_*\|_F \le K_1\kappa^{-1}\xi(p,\mathcal{G})\sqrt{\frac{\log p}{n}}\right) \ge P\left(\|B^+(S-\Omega_*^{-1})A^+\|_\infty \le \frac{K_1}{2\tau_1}\sqrt{\frac{\log p}{n}}\right) \ge 1 - C_1 p^{2-C_2K_1/2\tau_1},$$

which establishes (i) because $1 - C_1p^{2-C_2K_1/2\tau_1} \to 1$ as $K_1 \to \infty$. To establish (ii),

$$\begin{aligned}|A(\hat{\Omega}-\Omega_*)B|_1 &= |[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}}|_1 + |[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}^c}|_1 \\ &\le 4|[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}}|_1 \qquad (18)\\ &\le 4\xi(p,\mathcal{G})\|\hat{\Omega}-\Omega_*\|_F, \qquad (19)\end{aligned}$$

where (18) follows from Lemma 2 and (19) follows from the definition of $\xi(p,\mathcal{G})$.

Proof of Theorem 2

Lemma 4. *Let C_a and C_b be constants. Let a_n and b_n be sequences such that $|(\hat{A}_n - A)A^+|_1 \le C_a a_n$ and $|B^+(\hat{B}_n - B)|_1 \le C_b b_n$ with probability at least $1 - f(C_a)$ and $1 - g(C_b)$. Let $d_n = C_a a_n + C_b b_n + C_a C_b a_n b_n$. Then*

$$|\hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1 \ge |A(\Delta+\Omega_*)B|_1 - |[A\Omega_*B]_{\mathcal{G}}|_1 + d_n\left(|A\Delta B|_1 + 2|[A\Omega_*B]_{\mathcal{G}}|_1\right),$$

with probability at least $\min\{1-f(C_a), 1-g(C_b)\}$.

Proof of Lemma (4)**:** Let $|\hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1 \equiv V_1 - V_2$. First,

$$\begin{aligned}V_1 &= |\hat{A}_n(\Omega_*+\Delta)\hat{B}_n + A(\Omega_*+\Delta)B - A(\Omega_*+\Delta)B|_1 \\ &\ge |A(\Omega_*+\Delta)B|_1 - |A(\Omega_*+\Delta)B - \hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1, \qquad (20)\end{aligned}$$

by the triangle inequality. Also,

$$V_2 = |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_*B + A\Omega_*B|_1 \le |\hat{A}_n\Omega_*\hat{B}_n - [A\Omega_*B]_{\mathcal{G}}|_1 + |[A\Omega_*B]_{\mathcal{G}}|_1, \qquad (21)$$

so that from (20) and (21),

$$\begin{aligned}V_1 - V_2 \ge& |A(\Omega_*+\Delta)B|_1 - |[A\Omega_*B]_{\mathcal{G}}|_1 \\ &- |A(\Omega_*+\Delta)B - \hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_*B|_1. \qquad (22)\end{aligned}$$

Let $V_3 = -|A(\Omega_*+\Delta)B - \hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_*B|_1$. By a triangle inequality on the first term of V_3,

$$V_3 \ge -2|\hat{A}_n\Omega_*\hat{B}_n - A\Omega_*B|_1 - |\hat{A}_n\Delta\hat{B}_n - A\Delta B|_1. \qquad (23)$$

To bound (23), we need to bound functions of the form $|AMB - \hat{A}_n M \hat{B}_n|_1$ for arbitrary symmetric matrices M:

$$\begin{aligned}
&|AMB - \hat{A}_n M \hat{B}_n|_1 \\
&\qquad = |(A - \hat{A}_n)MB + AM(B - \hat{B}_n) + (A - \hat{A}_n)M(\hat{B}_n - B)|_1 \\
&\qquad \leq |(A - \hat{A}_n)MB|_1 + |AM(B - \hat{B}_n)|_1 + |(A - \hat{A}_n)M(\hat{B}_n - B)|_1. \qquad (24) \\
&\qquad = |(A - \hat{A}_n)A^+AMB|_1 + |AMBB^+(B - \hat{B}_n)|_1 \\
&\qquad\qquad + |(A - \hat{A}_n)A^+AMBB^+(\hat{B}_n - B)|_1, \qquad (25) \\
&\qquad \leq |AMB|_1 \Big\{ |(A - \hat{A}_n)A^+|_1 + |B^+(B - \hat{B}_n)|_1 \\
&\qquad\qquad + |(A - \hat{A}_n)A^+|_1 |B^+(B - \hat{B}_n)|_1 \Big\} \qquad (26) \\
&\qquad \leq |AMB|_1 \left(C_a a_n + C_b b_n + C_a C_b a_n b_n \right), \qquad (27)
\end{aligned}$$

where (24) follows from the triangle inequality; (25) follows from Assumption 2 and the definition of A^+ and B^+; (26) follows from the sub-multiplicative property of the $|\cdot|_1$ norm; and (27) holds with probability at least $\min\{1 - f(C_a), 1 - g(C_b)\}$. Applying (27) to both terms in (23) gives

$$V_3 \geq -2|\hat{A}_n \Omega_* \hat{B}_n - A\Omega_* B|_1 - |\hat{A}_n \Delta \hat{B}_n - A\Delta B|_1 \geq -d_n \left(2|[A\Omega_* B]_{\mathcal{G}}|_1 + |A\Delta B|_1 \right),$$

with probability at least $\min\{1 - f(C_a), 1 - g(C_b)\}$. Plugging this bound into (22) gives the result.

Lemma 5. *Suppose Assumptions 1–3 are true, $d_n = o(1)$, the bound in Lemma 4 holds, $\lambda \leq \epsilon\kappa(Q_\epsilon \tau)^{-1}$ for $\tau > 8$, where*

$$Q_\epsilon = \left\{ \left(\frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon} \right\}.$$

Then for all positive and sufficiently small ϵ, $\|B^+(S - \Omega_*^{-1})A^+\|_\infty \leq \lambda/2$, implies $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$.

Proof of Lemma 5: Let $\tilde{f}$ be the objective function from (11). Define $\tilde{D}(\Delta) = \tilde{f}(\Omega_* + \Delta) - \tilde{f}(\Omega_*)$ so that

$$\tilde{D}(\Delta) = \operatorname{tr}(S\Delta) + \log\det(\Omega_*) - \log\det(\Omega_* + \Delta) + \lambda_1 \left\{ |\hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n \Omega_* \hat{B}_n|_1 \right\}.$$

As in the proof of Lemma 1, we want to show that for $\Delta \in \mathcal{B}_\epsilon$, $\inf\{\tilde{D}(\Delta) : \|\Delta\|_F \leq \epsilon\} > 0$. Applying Lemma 4 to bound $|\hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n \Omega_* \hat{B}_n|_1$ and applying the same arguments

as in the proof of Lemma 1 to obtain (16),

$$\begin{aligned}
\tilde{D}(\Delta) \geq & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}\left(|\,[A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1\right) + \lambda\left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right) \\
& - \lambda d_n\left(|A\Delta B|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1\right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 - \lambda d_n\left(|A\Delta B|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1\right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 \\
& - \lambda d_n\left(|[A\Delta B]_{\mathcal{G}}|_1 + |[A\Delta B]_{\mathcal{G}^c}|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1\right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \left(\frac{3}{2} + d_n\right)\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 \\
& + \left(\frac{1}{2} - d_n\right)\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 - 2\lambda d_n|[A\Omega_* B]_{\mathcal{G}}|_1 \qquad (28)
\end{aligned}$$

and because $d_n = o(1)$ by assumption, for sufficiently large n, (28) implies

$$\begin{aligned}
\tilde{D}(\Delta) \geq & \frac{1}{8}\kappa\|\Delta\|_F^2 - \left(\frac{3}{2} + d_n\right)\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 - 2\lambda d_n|[A\Omega_* B]_{\mathcal{G}}|_1 \\
= & \|\Delta\|_F^2\left\{\frac{1}{8}\kappa - \left(\frac{3}{2} + d_n\right)\frac{\lambda}{\|\Delta\|_F}\xi(p,\mathcal{G}) - 2\lambda d_n\frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\|\Delta\|_F^2}\right\} \\
= & \|\Delta\|_F^2\left[\frac{1}{8}\kappa - \frac{\lambda}{\|\Delta\|_F}\left\{\left(\frac{3}{2} + d_n\right)\xi(p,\mathcal{G}) - 2d_n\frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\|\Delta\|_F}\right\}\right]. \qquad (29)
\end{aligned}$$

Since $\|\Delta\|_F = \epsilon$ and $\lambda \leq \epsilon\kappa(\tau Q_\epsilon)^{-1}$ for $\tau > 8$, where

$$Q_\epsilon = \left\{\left(\frac{3}{2} + d_n\right)\xi(p,\mathcal{G}) - 2d_n\frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon}\right\},$$

the inequality from (29) implies

$$D(\Delta) \geq \epsilon^2\left[\frac{1}{8}\kappa - \frac{\lambda}{\epsilon}\left\{\left(\frac{3}{2} + d_n\right)\xi(p,\mathcal{G}) - 2d_n\frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon}\right\}\right] \geq \epsilon^2\left(\frac{1}{8}\kappa - \frac{1}{\tau}\kappa\right) > 0,$$

which establishes the desired result.

Lemma 6. *If the conditions of Lemma 5 are true, then $\hat{\Delta} = \tilde{\Omega} - \Omega_*$ belongs to the set*

$$\left\{\Delta \in \mathbb{S}^{p_n} : |[A\Delta B]_{\mathcal{G}^c}|_1 \leq \frac{(3 + 2d_n)|[A\Delta B]_{\mathcal{G}}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1 - 2d_n}\right\}.$$

Proof of Lemma 6: Using the same arguments as in the proof of Lemma 1 from Negahban et al.

(2012), and from (28), we have

$$\begin{aligned}0 \leq D(\hat{\Delta}) \leq & -\frac{\lambda}{2}(1+2d_n)\left(|[A\Delta B]_{\mathcal{G}}|_1 + |[A\Delta B]_{\mathcal{G}^c}|_1\right) + \lambda\left(|[A\Delta B]_{\mathcal{G}^c}|_1 - |[A\Delta B]_{\mathcal{G}}|_1\right) \\ & - 2\lambda d_n |[A\Omega_* B]_{\mathcal{G}}|_1 \\ = & -\frac{\lambda}{2}\left\{(3+2d_n)|[A\Delta B]_{\mathcal{G}}|_1 - (1-2d_n)|[A\Delta B]_{\mathcal{G}^c}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1\right\}\end{aligned}$$

so that

$$|[A\Delta B]_{\mathcal{G}^c}|_1 \leq \frac{(3+2d_n)|[A\Delta B]_{\mathcal{G}}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n},$$

which is the desired inequality.

Proof of Theorem 2: Set $\lambda = K_2\tau_2^{-1}(n^{-1}\log p)^{1/2}$ and $\epsilon = \lambda Q_\epsilon \tau_2 \kappa^{-1}$. We can simplify the expression for ϵ by solving

$$\epsilon\kappa(K_2^2 n^{-1}\log p)^{-1/2} = \left\{\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G}) - 2d_n\frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon}\right\},$$

or equivalently,

$$\epsilon^2\kappa(K_2^2 n^{-1}\log p)^{-1/2} - \epsilon\left\{\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G})\right\} - 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1 = 0. \tag{30}$$

Using the quadratic formula to solve (30) for ϵ,

$$\begin{aligned}\epsilon = & \frac{K_2}{2\kappa}\sqrt{\frac{\log p}{n}}\left[\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G})\right. \\ & \left. + \left\{\left(\frac{3}{2}+d_n\right)^2\xi^2(p,\mathcal{G}) + \frac{16 d_n\kappa}{K_2}\sqrt{\frac{n}{\log p}}|[A\Omega_* B]_{\mathcal{G}}|_1\right\}^{1/2}\right].\end{aligned} \tag{31}$$

To simplify the result, we find an $\tilde{\epsilon}$ such that $\epsilon \leq \tilde{\epsilon}$. Then $\|\tilde{\Omega} - \Omega_*\|_F \leq \epsilon$ implies $\|\tilde{\Omega} - \Omega_*\|_F \leq \tilde{\epsilon}$, so $\|B^+(S-\Omega_*)A^+\|_\infty \leq \lambda/2$ also implies $\|\tilde{\Omega} - \Omega_*\|_F \leq \tilde{\epsilon}$. Viewing the square root in (31) as the Euclidean norm of the sum of the square root of its two terms, we use the triangle inequality to obtain

$$\epsilon \leq \frac{K_2}{\kappa}\sqrt{\frac{\log p}{n}}\left\{\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G}) + 2\left(\frac{d_n\kappa}{K_2}\sqrt{\frac{n}{\log p}}|[A\Omega_* B]_{\mathcal{G}}|_1\right)^{1/2}\right\} = \tilde{\epsilon}.$$

Then, applying Lemma 5 and Lemma A.3 from Bickel and Levina (2008), there exists constants C_3 and C_4 such that for sufficiently large n,

$$\begin{aligned}P\left(\|\hat{\Omega} - \Omega_*\|_F \leq \tilde{\epsilon}\right) & \geq P\left(\|B^+SA^+ - B^+\Omega_*^{-1}A^+\|_\infty \leq \frac{K_2}{2\tau_2}\sqrt{\frac{\log p}{n}}\right) \\ & \geq 1 - C_3 p^{2-C_4K_2/2\tau_2}\end{aligned}$$

which establishes (i) because $1 - C_3 p^{2-C_4K_2/2\tau_2} \to 1$ as $K_2 \to \infty$. To establish (ii), we bound

$|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1$. By the triangle inequality,

$$\begin{aligned}|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 &= |\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B + A\tilde{\Omega}B - A\Omega_* B|_1 \\ &\leq |\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B|_1 + |A\tilde{\Omega}B - A\Omega_* B|_1 \end{aligned} \tag{32}$$

and by the argument used to obtain the inequality in (27), $|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B|_1 \leq d_n|A\tilde{\Omega}B|_1$. Using this bound on the first term in (32),

$$|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 \leq d_n|A\tilde{\Omega}B|_1 + |A\tilde{\Omega}B - A\Omega_* B|_1. \tag{33}$$

Then, bounding the first term in (33), $|A\tilde{\Omega}B|_1 = |A\tilde{\Omega}B + A\Omega_* B - A\Omega_* B|_1 \leq |A\Omega_* B|_1 + |A\tilde{\Omega}B - A\Omega_* B|_1$ so that from (33),

$$|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 \leq d_n|A\Omega_* B|_1 + (d_n+1)|A\tilde{\Omega}B - A\Omega_* B|_1. \tag{34}$$

To bound the right term in the sum on the right hand side of (34), we apply Lemma 6 to $\tilde{\Delta} = \tilde{\Omega} - \Omega_*$

$$\begin{aligned}|A\tilde{\Omega}B - A\Omega_* B|_1 &\leq |[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + |[A\tilde{\Delta}B]_{\mathcal{G}^c}|_1 \\ &\leq |[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + \frac{(3+2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n} \\ &= \frac{(1-2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + (3+2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n} \\ &= \frac{4|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n}. \end{aligned} \tag{35}$$

Because $d_n = o(1)$, there exists constants C_5 and C_6 such that for some sufficiently large n, (35) implies $|A\tilde{\Omega}B - A\Omega_* B|_1 \leq C_5\|\tilde{\Omega} - \Omega_*\|_F\xi(p,\mathcal{G}) + C_6 d_n|[A\Omega_* B]_{\mathcal{G}}|_1$. Combining this with (34),

$$\begin{aligned}|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 &\leq (d_n+1)\left\{C_5\|\tilde{\Omega} - \Omega_*\|_F\xi(p,\mathcal{G}) + C_6 d_n|[A\Omega_* B]_{\mathcal{G}}|_1\right\} + d_n|[A\Omega_* B]_{\mathcal{G}}|_1. \\ &= C_5(d_n+1)\|\tilde{\Omega} - \Omega_*\|_F\xi(p,\mathcal{G}) + C_6(d_n^2 + d_n + d_n C_6^{-1})|[A\Omega_* B]_{\mathcal{G}}|_1,\end{aligned}$$

so that using $d_n = o(1)$ and the result from Theorem 2 (i) for $\|\tilde{\Omega} - \Omega_*\|_F$, we obtain the result.

References

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., et al. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The Journal of Molecular Diagnostics*, 8(1):51–61.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.

Chen, G. and Teboulle, M. (1994). A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1-3):81–101.

Chen, X., Xu, M., and Wu, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing*, 64(24):6459–6470.

Dalal, O. and Rajaratnam, B. (2014). G-AMA: sparse gaussian graphical model estimation via alternating minimization. *arXiv preprint arXiv:1405.3034*.

Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

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

Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics*, 11:599–608.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254—4278.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Mai, Q., Yang, Y., and Zou, H. (2015). Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99:29–42.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Negahban, S. N., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2015). Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2):439–454.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99:733–740.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J. and Forzani, L. (2014). On the existence of the weighted bridge penalized Gaussian likelihood precision matrix estimator. *Electronic Journal of Statistics*, 8:2693–2700.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Witten, D. M. and Tibshirani, R. J. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636.

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2015). Covariance-enhanced discriminant analysis. *Biometrika*, 102:33–45.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite 1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 93:19–35.