Electronic Journal of Statistics
Vol. 8 (2014) 2693–2700
ISSN: 1935-7524
DOI: 10.1214/14-EJS973

On the existence of the weighted bridge penalized Gaussian likelihood precision matrix estimator

Adam J. Rothman

School of Statistics
University of Minnesota
e-mail: arothman@umn.edu

and

Liliana Forzani

Instituto de Matemática Aplicada del Litoral (CONICET-UNL)
and Facultad de Ingeniería Química (UNL)
e-mail: liliana.forzani@gmail.com

Abstract: We establish a necessary and sufficient condition for the existence of the precision matrix estimator obtained by minimizing the negative Gaussian log-likelihood plus a weighted bridge penalty. This condition enables us to connect the literature on Gaussian graphical models to the literature on penalized Gaussian likelihood.

AMS 2000 subject classifications: Primary 62H12; secondary 62H20.
Keywords and phrases: High-dimensional data, precision matrix, ridge penalty, sparsity.

Received December 2014.

1. Introduction

Let $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^p$, and $\mathbb{R}^{p\times q}$ denote the sets of real numbers, positive real numbers, real valued p-dimensional vectors, and real valued p by q matrices, respectively. When $A, B \in \mathbb{R}^{p\times q}$, let A' be the transpose of A and let $A \circ B = [a_{ij}b_{ij}] \in \mathbb{R}^{p\times q}$ be the Hadamard product of A and B. When $A \in \mathbb{R}^{p\times p}$, let $\mathrm{tr}(A)$ denote the trace of A, let $|A|$ denote the determinant of A, and define $\mathbb{S}^p = \{A \in \mathbb{R}^{p\times p} : A = A'\}$. When $A \in \mathbb{S}^p$, let $\varphi_j(A), \varphi_{\min}(A)$, and $\varphi_{\max}(A)$ denote the jth largest eigenvalue, minimum eigenvalue, and maximum eigenvalue of A, respectively. Define $\mathbb{S}^p_0 = \{A \in \mathbb{S}^p : \varphi_{\min}(A) \geq 0\}$ and $\mathbb{S}^p_+ = \{A \in \mathbb{S}^p : \varphi_{\min}(A) > 0\}$.

Since the sample covariance matrix S is singular with probability one when the number of variables p is greater than the sample size n, regularization is required to estimate an inverse covariance matrix. There are many types of regularization that could be applied, e.g. a Moore–Penrose generalized inverse of S or the inverse of a linear combination of S and the identity matrix (Ledoit and Wolf, 2003). Pourahmadi (2011) reviews several approaches.

We study regularized inverse covariance estimators obtained by weighted bridge penalized Gaussian likelihood. The inverse covariance matrix estimator is

$$\hat{\Omega}_{\lambda,q}(M) = \arg\min_{\Omega \in \mathbb{S}^p_+} \left\{ \mathrm{tr}(\Omega S) - \log|\Omega| + \frac{\lambda}{q} \sum_{i,j} m_{ij} |\omega_{ij}|^q \right\}, \tag{1}$$

where $M \in \mathbb{S}^p$ is a known weight matrix with non-negative entries, λ is a non-negative tuning parameter, and $q \in [1,2]$ is the bridge parameter. The weight matrix M allows the user to incorporate prior information in the penalty, e.g. if it were known that the (i,j)th entry of the population inverse covariance matrix were non-zero or large, then one could set $m_{ij} = 0$. Without prior information on the location of small entries, one could set $M = M_{\mathrm{all}}$ or $M = M_{\mathrm{off}}$, where M_{all} is the weight matrix of ones and M_{off} is the weight matrix with ones on its off-diagonal and zeros on its diagonal.

The inverse covariance estimator obtained through (1) with $q = 1$ has been studied extensively. Yuan and Lin (2007) proposed and analyzed $\hat{\Omega}_{\lambda,1}(M_{\mathrm{off}})$ and Rothman et al. (2008), Lam and Fan (2009), and Ravikumar et al. (2011) established theoretical properties when $p > n$. Banerjee, El Ghaoui and d'Aspremont (2008) proposed and studied the optimization to compute $\hat{\Omega}_{\lambda,1}(M_{\mathrm{all}})$. The general weighted penalty in (1) with $q = 1$ was considered by Friedman, Hastie and Tibshirani (2008), Lu (2010), and Hsieh et al. (2012). The solution to (1) with $q = 1$, when it exists, could be computed with the graphical lasso algorithm (Yuan, 2008; Friedman, Hastie and Tibshirani, 2008) or the QUIC algorthm (Hsieh et al., 2011). Speed improvements to compute $\hat{\Omega}_{\lambda,1}(M_{\mathrm{all}})$ for large values of λ were proposed by Witten, Friedman and Simon (2011) and Mazumder and Hastie (2012). A benefit of using $q = 1$ is that for λ sufficiently large, a subset of the off-diagonal entries in $\hat{\Omega}_{\lambda,1}(M)$ are zero and when S is computed from n independent copies of a p-variate Normal random vector, the locations of non-zero entries in $\hat{\Omega}_{\lambda,1}(M)$ could be used to estimate the edges in a Gaussian graphical model (Yuan and Lin, 2007).

The use of the bridge penalty in (1) was proposed by Rothman et al. (2008) in the special case when $M = M_{\mathrm{off}}$. They proposed an algorithm to compute $\hat{\Omega}_{\lambda,q}(M_{\mathrm{off}})$ for $q \in [1,2]$, but they only explored its use when $q = 1$. Witten and Tibshirani (2009) derived a closed-form solution to compute $\hat{\Omega}_{\lambda,2}(M_{\mathrm{all}})$.

The inverse covariance estimator $\hat{\Omega}_{\lambda,q}(M)$ could be used within methods for supervised learning, e.g. discriminant analysis (Rothman et al., 2008), covariance regularized regression (Witten and Tibshirani, 2009), and multivariate regression with covariance estimation (Rothman, Levina and Zhu, 2010b). When the population inverse covariance matrix is sparse or approximately sparse, there are theoretical advantages to set $q = 1$, but in general settings $q \in (1,2]$ may be useful.

There has been some study of sufficient conditions for the solution to (1) to exist: Banerjee, El Ghaoui and d'Aspremont (2008) showed that the solution to (1) exists with probability one if $q = 1$, $M = M_{\mathrm{all}}$, and $\lambda > 0$; Ravikumar et al. (2011) showed that the solution to (1) exists if $q = 1$, $M = M_{\mathrm{off}}$, $\lambda > 0$, and $S \circ I \in \mathbb{S}_+^p$; and Lu (2010) showed that the solution to (1) exists if $q = 1$ and $S + \lambda M \circ I \in \mathbb{S}_+^p$. Necessary and sufficient conditions for there to exist a solution to (1) are still unknown, and we establish these in this article.

2. Properties of the optimization

2.1. Solution existence

The optimization in (1) is strictly convex. When an optimal solution exists, this solution is the unique global minimizer. Let $\mathcal{U} = \{(i,j) : m_{ij} = 0\}$ be the set of unpenalized indices.

Theorem 1. *Suppose that $\lambda \in \mathbb{R}_+$ and $q \in (1, 2]$. An optimal solution to (1) exists if and only if the set $\mathcal{A}(M) = \{\Sigma \in \mathbb{S}_+^p : \sigma_{ij} = s_{ij} \text{ when } (i,j) \in \mathcal{U}\}$ is not empty.*

Theorem 2. *Suppose that $q = 1$. An optimal solution to (1) exists if and only if the set $\mathcal{A}_1(\lambda, M) = \{\Sigma \in \mathbb{S}_+^p : |\sigma_{ij} - s_{ij}| \leq \lambda m_{ij}$ for all $i, j\}$ is not empty.*

We prove Theorems 1 and 2 in Appendix A. Hsieh et al. (2012) stated a dual problem to (1) with $q = 1$ for which $\mathcal{A}_1(\lambda, M)$ is the feasible set; however, we do not use this duality to prove Theorem 2. Our proof technique is also used to prove Theorem 1.

When the diagonal entries are unpenalized, i.e. $\max_j m_{jj} = 0$, the condition that $\mathcal{A}(M)$ is not empty is necessary and sufficient for the maximum likelihood estimator of the inverse covariance matrix to exist in the Gaussian graphical model with edge set $\mathcal{U}$; see, for example, Theorem 2.1 of Uhler (2012). As a consequence, we have the following example.

Example 1. Suppose that S is computed from an iid sample of size n from a p-variate normal distribution with positive definite covariance matrix and $m_{ij} = 1(|i - j| > k)$, where $k \in \{0, \ldots, p-1\}$. Then the set of unpenalized indices $\mathcal{U}$ is the edge set of a Chordal graph with maximum clique size $k + 1$. From Corollary 2.3 of Uhler (2012), the MLE of the Gaussian graphical model with edge set $\mathcal{U}$ exists with probability one if and only if $k < n - 1$. This result is also in Buhl (1993) and the MLE in this case can be computed by inverse Cholesky banding (Rothman, Levina and Zhu, 2010a). Thus, by Theorem 1, the solution to (1) with $q \in (1, 2]$ exists with probability one if and only if $k < n-1$.

Define the weighted soft-thresholded sample covariance matrix by $\text{soft}(S, \lambda M) = [\text{sign}(s_{ij})(|s_{ij}| - \lambda m_{ij})_+]$, where $(y)_+ = y1(y > 0)$.

Corollary 1. *An optimal solution to (1) exists if at least one of the following conditions is true: (i) $\lambda \in \mathbb{R}_+$ and $\min_j m_{jj} > 0$; (ii) $S \circ I \in \mathbb{S}_+^p$, $\lambda \in \mathbb{R}_+$, and $\min_{i \neq j} m_{ij} > 0$; (iii) $S \in \mathbb{S}_+^p$; (iv) $\text{soft}(S, \lambda M) \in \mathbb{S}_+^p$.*

We prove Corollary 1 in Appendix A. In particular, if all diagonal entries are penalized in (1), then a solution exists for any choice of the off-diagonal of M. Also, if all off-diagonal entries are penalized and the diagonal entries of

S are positive, then a solution exists for any choice of the diagonal of M, e.g. the diagonal of zeros. One could check if $\text{soft}(S, \lambda M)$ is positive definite, which guarantees an optimal solution to (1) exists, but we have encountered several examples where the realization of $\text{soft}(S, \lambda M)$ was indefinite and an optimal solution to (1) with $q = 1$ existed.

Since $\mathcal{A}_1(\lambda, M) \subset \mathcal{A}(M)$ for all $\lambda \geq 0$, we see that if $\hat{\Omega}_{\lambda,1}(M)$ exists for some $\lambda \in \mathbb{R}_+$, then $\hat{\Omega}_{\hat{\lambda},q}(M)$ exists for all $(\hat{\lambda}, q) \in \mathbb{R}_+ \times (1, 2]$. Also, if there exists a $\bar{\Sigma} \in \mathcal{A}(M)$, then there exists a $\bar{\lambda} \in \mathbb{R}_+$ sufficiently large such that $\bar{\Sigma} \in \mathcal{A}_1(\bar{\lambda}, M)$. From this and the definition of $\mathcal{A}(M)$ we have that if $\hat{\Omega}_{\lambda,q}(M)$ exists for some $(\lambda, q) \in \mathbb{R}_+ \times (1, 2]$, then $\hat{\Omega}_{\tilde{\lambda},\tilde{q}}(M)$ exists for all $(\tilde{\lambda}, \tilde{q}) \in \mathbb{R}_+ \times (1, 2]$ and there exists a $\bar{\lambda} \in \mathbb{R}_+$ sufficiently large such that $\hat{\Omega}_{\bar{\lambda},1}(M)$ exists.

2.2. *The effect of unpenalized entries*

From first-order conditions and convexity, a feasible point $\hat{\Omega} \in \mathbb{S}^p_+$ is optimal for (1) if and only if it solves the zero gradient equation when $q \in (1, 2]$ or the zero subgradient equation when $q = 1$. In both cases, this equation can be written as

$$0 = S - \hat{\Omega}^{-1} + \lambda M \circ G, \tag{2}$$

where the (i, j)th entry of G is a function evaluated at the (i, j)th entry of $\hat{\Omega}$, e.g. Meinshausen (2008) gave an expression for (2) when $q = 1$ and $M = M_{\text{off}}$. We see that if $\hat{\Omega}$ is optimal for (1) then $(\hat{\Omega}^{-1})_{ij} = s_{ij}$ when $(i, j) \in \mathcal{U}$.

3. Discussion

In an earlier submission of this work, we proposed an accelerated majorize minimize algorithm to solve (1) with $q = 2$ and any weight matrix M. This special case can be used to iteratively solve (1) with $q \in [1, 2)$ (Rothman et al., 2008). In our numerical experiments, our algorithm was competitive for all values of the tuning parameter when $M = M_{\text{off}}$, but it was uncompetitive for mid-range values of the tuning parameter when $m_{ij} = 1(|i - j| > k)$ with $k \geq 1$. We are currently addressing this problem.

Acknowledgments

The authors thank the editor and referees for helpful suggestions. The authors thank Xiaotong Shen for pointing out connections to Gaussian graphical models and thank Arindam Banerjee for a helpful discussion. Rothman's research is supported in part by the National Science Foundation DMS-1105650.

Appendix A: Proofs

For $c, d \in \mathbb{R}_+$, define the set $\mathcal{G}_{c,d} = \{\Omega \in \mathbb{S}^p_+ : c \leq \varphi_{\min}(\Omega) \leq \varphi_{\max}(\Omega) \leq d\}$.

To prove Theorems 1 and 2 we will use the following Lemma, which is a modification of an argument made by Bien and Tibshirani (2011).

Lemma 1. *Suppose that a function $v : \mathbb{S}_+^p \to \mathbb{R}$ satisfies $v(\Omega) \geq a + b\ \mathrm{tr}(\Omega) - \log|\Omega|$ for all $\Omega \in \mathbb{S}_+^p$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}_+$. If $\Omega_0 \in \mathbb{S}_+^p$, then there exist $c, d \in \mathbb{R}_+$ such that $\{\Omega \in \mathbb{S}_+^p : v(\Omega) \leq v(\Omega_0)\} \subset \mathcal{G}_{c,d}$.*

Proof of Lemma 1. Our proof modifies an argument of Bien and Tibshirani (2011). We have that

$$v(\Omega) \geq a + \sum_{j=1}^{p} \{b\varphi_j(\Omega) - \log \varphi_j(\Omega)\}. \tag{3}$$

Define $r(;b) : \mathbb{R}_+ \to \mathbb{R}$ by $r(x;b) = xb - \log x$. Then $\nabla_x r(x;b) = b - 1/x$, so $\hat{x} = 1/b$ minimizes $r(;b)$ and $r(x;b) \geq r(\hat{x};b) = 1 + \log b$. From (3),

$$\begin{aligned} v(\Omega) &\geq a + \sum_{j=1}^{p} r\{\varphi_j(\Omega); b\} \\ &= a + r\{\varphi_j(\Omega); b\} + \sum_{k=1, k\neq j}^{p} r\{\varphi_k(\Omega); b\} \\ &\geq a + r\{\varphi_j(\Omega); b\} + (p-1)(1 + \log b), \end{aligned}$$

for $j \in \{1, \ldots, p\}$. Given any $\Omega_0 \in \mathbb{S}_+^p$ define $\mathcal{R}(\Omega_0) = \{\Omega \in \mathbb{S}_+^p : v(\Omega) \leq v(\Omega_0)\}$. For all $\bar{\Omega} \in \mathcal{R}(\Omega_0)$, $v(\Omega_0) \geq v(\bar{\Omega}) \geq a + r\{\varphi_j(\bar{\Omega}); b\} + (p-1)(1 + \log b)$ and so for all $\bar{\Omega} \in \mathcal{R}(\Omega_0)$,

$$\varphi_j(\bar{\Omega}) \in \{\phi \in \mathbb{R}_+ : r(\phi; b) \leq \delta_0\}, \quad j = 1, \ldots, p,$$

where $\delta_0 = v(\Omega_0) - (p-1)(1 + \log b) - a$. Since $r(;b)$ is decreasing on $(0, \hat{x})$, increasing on $(\hat{x}, \infty)$, $\lim_{x\to 0^+} r(x;b) = \infty$, and $\lim_{x\to\infty} r(x;b) = \infty$, there exist $c, d \in \mathbb{R}_+$ such that $c \leq \varphi_{\min}(\bar{\Omega}) \leq \varphi_{\max}(\bar{\Omega}) \leq d$ for all $\bar{\Omega} \in \mathcal{R}(\Omega_0)$. □

Proof of Theorem 1. Let f be the objective function in (1). Suppose that $\mathcal{A}(M)$ is not empty and that $\bar{\Sigma} \in \mathcal{A}(M)$. Then

$$\begin{aligned} f(\Omega) + \log|\Omega| &= \mathrm{tr}\{\Omega\bar{\Sigma}\} + \mathrm{tr}\{\Omega(S - \bar{\Sigma})\} + \frac{\lambda}{q} \sum_{i,j} m_{ij}|\omega_{ij}|^q \\ &= a_1 + a_2 + a_3. \end{aligned}$$

We have that $a_1 \geq \varphi_{\min}(\bar{\Sigma}) \sum_{j=1}^{p} \varphi_j(\Omega)$. Also,

$$a_2 + a_3 = \sum_{i,j} \{\omega_{ij}(s_{ij} - \bar{\sigma}_{ij}) + \lambda m_{ij}|\omega_{ij}|^q/q\}.$$

If $m_{ij} = 0$ then $s_{ij} = \bar{\sigma}_{ij}$ so the (i,j)th entry in this sum is zero. The non-zero entries in this sum are also bounded from below because $\lambda m_{ij}|\omega_{ij}|^q/q > |\omega_{ij}(s_{ij} - \bar{\sigma}_{ij})|$ if $|\omega_{ij}|$ is sufficiently large. Thus there exists an $a \in \mathbb{R}$ for which $a_2 + a_2 \geq a$. We have shown that $f(\Omega) \geq a + b\mathrm{tr}(\Omega) - \log|\Omega|$, where $b = \varphi_{\min}(\bar{\Sigma}) \in \mathbb{R}_+$. So for any $\Omega_0 \in \mathbb{S}_+^p$, Lemma 1 implies that there exists $c, d \in \mathbb{R}_+$ such that $\{\Omega \in \mathbb{S}_+^p : f(\Omega) \leq f(\Omega_0)\} \subset \mathcal{G}_{c,d}$. Given $\Omega_0 \in \mathbb{S}_+^p$, we can shrink the feasible set in (1) to the compact subset $\mathcal{G}_{c,d}$ and since f is convex, a global minimizer exists in $\mathcal{G}_{c,d}$.

Now suppose that a global minimizer $\hat{\Sigma}^{-1} \in \mathbb{S}_+^p$ of (1) exists. Then $\hat{\Sigma} \in \mathbb{S}_+^p$ and since $\nabla f(\hat{\Sigma}^{-1}) = 0$, we have that $S - \hat{\Sigma} + \lambda M \circ G = 0$, where the (i,j)th entry of G is determined by the (i,j)th entry of $\hat{\Sigma}^{-1}$, so $s_{ij} = \hat{\sigma}_{ij}$ when $(i,j) \in \mathcal{U}$. Thus $\hat{\Sigma} \in \mathcal{A}(M)$, so $\mathcal{A}(M)$ is not empty. ☐

Proof of Theorem 2. Let u be the objective function in (1) with $q = 1$. Suppose that $\mathcal{A}_1(\lambda, M)$ is not empty and take $\bar{\Sigma} \in \mathcal{A}_1(\lambda, M)$. Then

$$u(\Omega) + \log|\Omega| = \mathrm{tr}\{\Omega\bar{\Sigma}\} + \mathrm{tr}\{\Omega(S - \bar{\Sigma}\} + \lambda \sum_{i,j} m_{ij}|\omega_{ij}|$$

$$= a_1 + a_2 + a_3.$$

We have that $a_1 \geq \varphi_{\min}(\bar{\Sigma}) \sum_{j=1}^{p} \varphi_j(\Omega)$. Also,

$$a_2 + a_3 = \sum_{i,j} \{\omega_{ij}(s_{ij} - \bar{\sigma}_{ij}) + \lambda m_{ij}|\omega_{ij}|\}.$$

Since $\bar{\Sigma} \in \mathcal{A}_1(\lambda, M)$, $|s_{ij} - \bar{\sigma}_{ij}| \leq \lambda m_{ij}$ for all i, j, so $a_2 + a_3 \geq 0$. Therefore, $u(\Omega) \geq a + b\mathrm{tr}(\Omega) - \log|\Omega|$ for all $\Omega \in \mathbb{S}_+^p$, where $a = 0$ and $b = \varphi_{\min}(\bar{\Sigma})$. Thus, for any $\Omega_0 \in \mathbb{S}_+^p$, Lemma 1 implies that there exists $c, d \in \mathbb{R}_+$ such that $\{\Omega \in \mathbb{S}_+^p : u(\Omega) \leq u(\Omega_0)\} \subset \mathcal{G}_{c,d}$. Given $\Omega_0 \in \mathbb{S}_+^p$, we can shrink the feasible set in (1) to the compact subset $\mathcal{G}_{c,d}$ and since u is convex, a global minimizer exists in $\mathcal{G}_{c,d}$.

Now suppose that a global minimizer $\hat{\Sigma}^{-1} \in \mathbb{S}_+^p$ of (1) exists. From the discussion in Section 2.2, $S - \hat{\Sigma} + \lambda M \circ G = 0$, where, in this case, G has entries in $[-1, 1]$. So $\hat{\Sigma} - S = \lambda M \circ G$, which implies that $|\hat{\sigma}_{ij} - s_{ij}| = \lambda m_{ij}|g_{ij}| \leq \lambda m_{ij}$ for all i, j. Thus $\hat{\Sigma} \in \mathcal{A}_1(\lambda, M)$. ☐

Proof of Corollary 1. Since $\mathcal{A}_1(\lambda, M) \subset \mathcal{A}(M)$ for all $\lambda \geq 0$, by Theorems 1 and 2, we only need to show that each of the four conditions imply that $\mathcal{A}_1(\lambda, M)$ is not empty.

For (i), take $\bar{\Sigma}_1 = S + \lambda \min_j m_{jj} I$. Then $\bar{\Sigma}_1 \in \mathcal{A}_1(\lambda, M)$. For (ii), pick an $\alpha \in \mathbb{R}_+$ sufficiently close to zero and let $\bar{\Sigma}_2 = (1-\alpha)S + \alpha S \circ I$. Then $\bar{\Sigma}_2 \in \mathbb{S}_+^p$, $\bar{\sigma}_{2,jj} = s_{jj}$, and when $i \neq j$, $|\bar{\sigma}_{2,ij} - s_{ij}| = \alpha|s_{ij}| \leq \lambda m_{ij}$ provided that α is sufficiently close to zero. So $\bar{\Sigma}_2 \in \mathcal{A}_1(\lambda, M)$. For (iii), $S \in \mathcal{A}_1(\lambda, M)$. For (iv) take $\bar{\Sigma}_4 = \mathrm{soft}(S, \lambda M)$. From Rothman, Levina and Zhu (2009), $|\bar{\sigma}_{4,ij} - s_{ij}| \leq \lambda m_{ij}$ so $\bar{\Sigma}_4 \in \mathcal{A}_1(\lambda, M)$. ☐

References

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research* **9** 485–516. MR2417243

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics* **20** 263–270. MR1241392

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

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. and Ravikumar, P. K. (2011). Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation. In *Advances in Neural Information Processing Systems*, **24** 2330–2338. MIT Press, Cambridge, MA.

Hsieh, C.-J., Banerjee, A., Dhillon, I. S. and Ravikumar, P. K. (2012). A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems*, **25** (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 2330–2338. Curran Associates, Inc.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* **37** 4254–4278. MR2572459

Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88** 365–411. MR2026339

Lu, Z. (2010). Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Its Applications* **31** 2000–2016. MR2678953

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale Graphical Lasso. *Journal of Machine Learning Research* **13** 781–794. MR2913718

Meinshausen, N. (2008). A note on the lasso for graphical Gaussian model selection. *Statistics and Probability Letters* **78** 880–884. MR2398362

Pourahmadi, M. (2011). Modeling covariance matrices: The GLM and regularization perspectives. *Statistical Science* **26** 369–387. MR2917961

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980. MR2836766

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104** 177–186. MR2504372

Rothman, A. J., Levina, E. and Zhu, J. (2010a). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. MR2672482

Rothman, A. J., Levina, E. and Zhu, J. (2010b). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19** 947–962. MR2791263

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515. MR2417391

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics* **40** 238–261. MR3014306

Witten, D. M., Friedman, J. H. and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20** 892–900. MR2878953

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of The Royal Statistical Society Series B* **71** 615–636. MR2749910

Yuan, M. (2008). Efficient computation of L1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics* **17** 809–826. MR2649068

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824