

Properties of optimizations used in penalized Gaussian likelihood inverse covariance matrix estimation

Adam J. Rothman
School of Statistics
University of Minnesota

October 8, 2014, joint work with Liliana Forzani

Estimating large covariance matrices and their inverses

Two different problems

A. Estimating Σ_*

B. Estimating Σ_*^{-1}

Desirable properties

- low computational cost
- works well in applications, e.g. classification
- exploits variable ordering when appropriate

An excellent book on the topic:

M. Pourahmadi (2013) *High-Dimensional Covariance Estimation*. Wiley, New York.

Estimating the covariance matrix Σ_*

General purpose methods

- Shrinking the eigenvalues of S (Haff, 1980; Dey and Srinivasan, 1985; Ledoit and Wolf, 2003).
- Element-wise thresholding of S (Bickel and Levina, 2008; El Karoui, 2008; Rothman, Levina, and Zhu, 2009; Cai & Zhou 2010; Cai & Liu, 2011).
- Matrix-logarithm reparameterization (Deng & Tsui, 2010)
- lasso-penalized Gaussian likelihood (Lam & Fan, 2009; Bien & Tibshirani, 2011).
- Other sparse and positive definite methods (Rothman, 2012; Xue, Ma, & Zou, 2012; Liu, Wang, & Zhao, 2013).
- Σ_* with reduced effective rank (Bunea & Xiao, 2012).
- Elliptical copula correlation matrix modeled as a low rank plus diagonal matrix (Wegkamp & Zhao, 2013).
- Approximate factor model with sparse error covariance (Fan, Liao & Minchev, 2013)

Estimating the inverse covariance matrix Σ_*^{-1}

General purpose methods

- Eigenvalue shrinkage (Ledoit and Wolf, 2003).
- Bayesian methods (Wong et al., 2003; Dobra et al, 2004).
- Penalized likelihood [References given soon.](#)

Estimating Σ_*^{-1} or Σ_* in applications

- Classification (Bickel & Levina, 2004; Gaynanova, Booth, & Wells, 2014)
- Regression (Witten & Tibshirani, 2009; Cook, Forzani, & Rothman, 2013)
- Multiple output regression (Rothman, Levina, & Zhu, 2010)
- Sufficient dimension reduction (Cook, Forzani, & Rothman, 2012)

Estimating Σ_*^{-1} and Gaussian graphical models

- $(X_1, \dots, X_p)' \sim N_p(0, \Sigma_*)$, $\Sigma_* \in \mathbb{S}_+^p$.
- The undirected graph $G = (V, E)$ has
 - vertex set $V = \{1, \dots, p\}$ and
 - edge set $E = \{(i, j) : (\Sigma_*^{-1})_{ij} \neq 0\}$.
- Selection (Drton & Perlman, 2004; Kalisch & Bühlmann, 2007; Meinshausen and Bühlmann 2006).
- Simultaneous estimation and selection (Yuan & Lin, 2007; Yuan, 2010; Cai, Liu, & Luo, 2011; Ren, Sun, Zhang, & Zhou, 2013).
- Estimation when E is known (Dempster, 1972; Buhl, 1993; Drton et al., 2009; Uhler, 2012).

Part I: On solution existence in penalized
Gaussian likelihood estimation of Σ_*^{-1}

Unpenalized Gaussian likelihood estimation of Σ_*^{-1}

S is from an iid sample of size n from $N_p(\mu_*, \Sigma_*)$, where $\Sigma_* \in \mathbb{S}_+^p$. The optimization for the MLE of Σ_*^{-1} is

$$\hat{\Omega} = \arg \min_{\Omega \in \mathbb{S}_+^p} \{\text{tr}(\Omega S) - \log \det(\Omega)\}$$

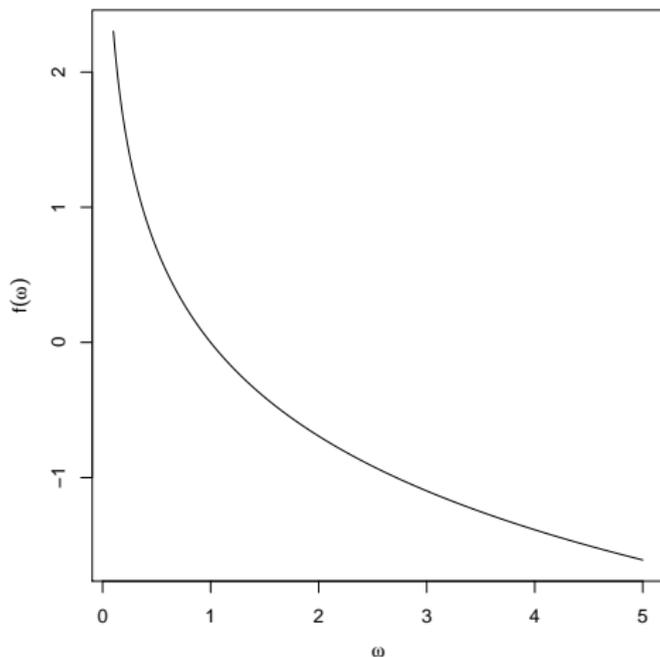
Set the gradient at $\hat{\Omega}$ to zero:

$$S - \hat{\Omega}^{-1} = 0$$

- The solution, when it exists, is $\hat{\Omega} = S^{-1}$.
- $\hat{\Omega}$ exists (with probability one) if and only if $n > p$ (Dykstra, 1970).

Illustration – No solution when $n \leq p$

If $n = p = 1$, then $f(\omega) = 0\omega - \log \omega$.



Lasso penalized likelihood estimation of Σ_*^{-1}

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

- $\hat{\Omega}_{\lambda,1}(M_{\text{off}})$ Yuan & Lin (2007), Rothman et al. (2008), Lam & Fan (2009), and Ravikumar, Wainwright, Raskutti, and Yu (2008).
- $\hat{\Omega}_{\lambda,1}(M_{\text{all}})$ Banerjee et al. (2008), Friedman et al. (2008).
- Algorithms to compute $\hat{\Omega}_{\lambda,1}(M)$ for some M Yuan (2008)/Friedman et al. (2008), Lu (2008), and Hsieh, Dhillon, Ravikumar, & A. Banerjee (2012).

Review of work on solution existence

- Banerjee et al. (2008) showed that a solution exists if $M = M_{\text{all}}$ and $\lambda > 0$.
- Ravikumar et al. (2008) showed that a solution exists if $M = M_{\text{off}}$, $\lambda > 0$, and $S \circ I \in \mathbb{S}_+^p$.
- Lu (2008) showed that a solution exists if $S + \lambda M \circ I \in \mathbb{S}_+^p$.

Lasso-penalized likelihood solution existence

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

Theorem. The solution $\hat{\Omega}_{\lambda,1}(M)$ exists if and only if $A_{1,\lambda}(M) = \{\Sigma \in \mathbb{S}_+^p : |\sigma_{ij} - s_{ij}| \leq m_{ij}\lambda\}$ is not empty.

$A_{1,\lambda}(M)$ is the feasible set in the dual problem (Hsieh, Dhillon, Ravikumar, & A. Banerjee, 2012), but our proof technique is more general.

Corollary. $\hat{\Omega}_{\lambda,1}(M)$ exists if at least one of the following holds:

- 1 $\lambda > 0$ and $\min_j m_{jj} > 0$;
- 2 $\lambda > 0$, $S \circ I \in \mathbb{S}_+^p$ and $\min_{i \neq j} m_{ij} > 0$;
- 3 $S \in \mathbb{S}_+^p$;
- 4 $\text{soft}(S, \lambda M) \in \mathbb{S}_+^p$.

Entry agreement between $\hat{\Omega}_{\lambda,1}(M)^{-1}$ and S

Define $\mathcal{U} = \{(i, j) : m_{ij} = 0\}$.

Remark. When $\hat{\Omega}_{\lambda,1}(M)$ exists,

$$\{\hat{\Omega}_{\lambda,1}(M)^{-1}\}_{ij} = s_{ij} \quad \text{for } (i, j) \in \mathcal{U}.$$

Justification. The zero subgradient equation is

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

Bridge penalized likelihood estimation of Σ_*^{-1}

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

- $q \in (1, 2]$.
- Why use the Bridge penalty?
- Rothman et al. (2008) developed an algorithm to compute $\hat{\Omega}_{\lambda,q}(M_{\text{off}})$.
- Witten and Tibshirani (2009) developed a closed-form solution to compute $\hat{\Omega}_{\lambda,q=2}(M_{\text{all}})$.

Bridge-penalized likelihood solution existence

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

Recall that $\mathcal{U} = \{(i, j) : m_{ij} = 0\}$.

Theorem. Suppose that $\lambda > 0$ and $q \in (1, 2]$. The solution $\hat{\Omega}_{\lambda,q}(M)$ exists if and only if $A_q(M) = \{\Sigma \in \mathbb{S}_+^p : \sigma_{ij} = s_{ij} \text{ when } (i, j) \in \mathcal{U}\}$ is not empty.

Corollary. Suppose $\lambda > 0$ and $q \in (1, 2]$. $\hat{\Omega}_{\lambda,q}(M)$ exists if and only if the MLE of the Gaussian graphical model with edge set \mathcal{U} exists.

Example. $m_{ij} = 1(|i - j| > k)$. Then \mathcal{U} is an edge-set for a Chordal graph with maximum clique size $k + 1$. So $\hat{\Omega}_{\lambda,q}(M)$ exists if and only if $n \geq k + 2$.

Entry agreement between $\hat{\Omega}_{\lambda,q}(M)^{-1}$ and S

Recall $\mathcal{U} = \{(i, j) : m_{ij} = 0\}$.

Remark. When $\hat{\Omega}_{\lambda,q}(M)$ exists,

$$\{\hat{\Omega}_{\lambda,q}(M)^{-1}\}_{ij} = s_{ij} \quad \text{for } (i, j) \in \mathcal{U}.$$

Bridge and lasso connections

Recall that $\mathcal{U} = \{(i, j) : m_{ij} = 0\}$

$$A_{1,\lambda}(M) = \{\Sigma \in \mathbb{S}_+^p : |\sigma_{ij} - s_{ij}| \leq m_{ij}\lambda\}$$

$$A_q(M) = \{\Sigma \in \mathbb{S}_+^p : \sigma_{ij} = s_{ij} \text{ when } (i, j) \in \mathcal{U}\}$$

- $A_{1,\lambda}(M) \subset A_q(M)$.
- If $\hat{\Omega}_{\lambda,1}(M)$ exists, then $\hat{\Omega}_{\tilde{\lambda},q}(M)$ exists for all $(\tilde{\lambda}, q) \in (0, \infty) \times (1, 2]$.
- If $\hat{\Omega}_{\lambda,q}(M)$ exists for some $(\lambda, q) \in (0, \infty) \times (1, 2]$, then it exists for all $(\lambda, q) \in (0, \infty) \times (1, 2]$, and there exists a $\bar{\lambda}$ sufficiently large so that $\hat{\Omega}_{\bar{\lambda},1}(M)$ exists.

Part II: new algorithm for the ridge penalty

The SPICE algorithm to compute the Bridge solution

Rothman, Bickel, Levina, & Zhu (2008)'s iterative algorithm to minimize

$$\text{tr}(\Omega S) - \log \det(\Omega) + \frac{\lambda}{q} \sum_{i \neq j} |\omega_{ij}|^q$$

- Each iteration minimizes

$$\text{tr}(\Omega S) - \log |\Omega| + \frac{\lambda}{2} \sum_{i \neq j} m_{ij} \omega_{ij}^2.$$

- Re-parameterize using Cholesky factorization: $\Omega = T'T$.
- Minimize with cyclical coordinate descent.
- Computational complexity $O(p^3)$.

Accelerated MM algorithm to compute the Ridge solution (our proposal)

Minimizes the objective function f ,

$$f(\Omega) = \text{tr}(\Omega S) - \log \det(\Omega) + 0.5\lambda \sum_{i,j} m_{ij} \omega_{ij}^2.$$

- At the k th iteration, the next iterate Ω_{k+1} is the minimizer of a majorizing function to f at Ω_k .
- Every few iterations a minorizing function to f at Ω_k is minimized. This minimizer is accepted if it improves objective function value.
- Computational complexity $O(p^3)$.

Majorizing f at Ω_k part 1

Decompose the penalty:

$$\sum_{i,j} m_{ij} \omega_{ij}^2 = \max_{i,j} m_{ij} \sum_{i,j} \omega_{ij}^2 - \sum_{i,j} (\max_{i,j} m_{ij} - m_{ij}) \omega_{ij}^2.$$

Replace the second term with a linear approximation at our current iterate Ω_k to get the majorizer.

Majorizing f at Ω_k part 2

- The minimizer of the majorizer to f at Ω_k is

$$\Omega_{k+1} = \arg \min_{\Omega \in \mathbb{S}_+^p} \left\{ \text{tr}(\Omega \tilde{S}) - \log \det(\Omega) + 0.5 \tilde{\lambda} \sum_{i,j} \omega_{ij}^2 \right\},$$

where $\tilde{S} = S - \lambda \Omega_k \circ [\max_{i,j} m_{ij} - m_{ij}]$ and $\tilde{\lambda} = \lambda \max_{i,j} m_{ij}$.

- Closed-form solution derived by [Witten and Tibshirani \(2009\)](#).

Acceleration by minorizing f at Ω_k

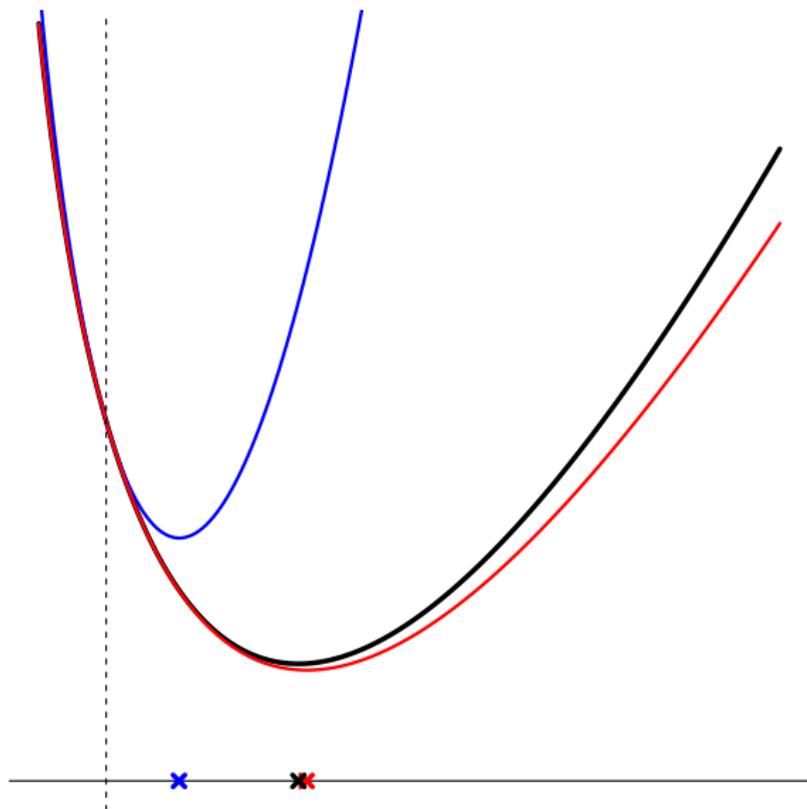
- Get the minorizer by replacing the entire penalty $\sum_{i,j} m_{ij} \omega_{ij}^2$ with a linear approximation at our current iterate Ω_k .

- The minimizer of the minorizer (when it exists) is

$$\tilde{\Omega}_{k+1} = (S + \lambda \Omega_k \circ M)^{-1}$$

- Accept $\tilde{\Omega}_{k+1}$ if $f(\tilde{\Omega}_{k+1}) < f(\Omega_k)$.

Illustration



Algorithm convergence

Theorem. Suppose that acceleration attempts are stopped after a finite number of iterations and the algorithm is initialized at $\Omega_0 \in \mathbb{S}_+^p$. If the global minimizer $\hat{\Omega}_{\lambda,2}(M)$ exists, then

$$\lim_{k \rightarrow \infty} \|\Omega_k - \hat{\Omega}_{\lambda,2}(M)\| = 0.$$

Also, if the algorithm converges, then it converges to the global minimizer.

ADMM algorithm

We can modify graphical lasso via ADMM (Boyd, 2011; Scheinberg et al., 2010) to minimize

$$\text{tr}(\Omega S) - \log \det(\Omega) + \frac{\lambda}{2} \sum_{i,j} m_{ij} z_{ij}^2$$

subject to $\Omega = Z$. Pick starting values and some $\rho > 0$ and repeat 1–3 until convergence

1. $\Omega_{k+1} = \arg \min_{\Omega \in \mathbb{S}_+^p} \left\{ \text{tr}(\Omega \tilde{S}) - \log |\Omega| + 0.5\rho \sum_{i,j} \omega_{ij}^2 \right\}$
where $\tilde{S} = S + \rho(U_k - Z_k)$.
2. $Z_{k+1} = (\Omega_{k+1} + U_k) \circ Q$, where $q_{ij} = (1 + \lambda m_{ij} / \rho)^{-1}$.
3. $U_{k+1} = U_k + \Omega_{k+1} - Z_{k+1}$.

Simulation: our algorithm vs the SPICE algorithm

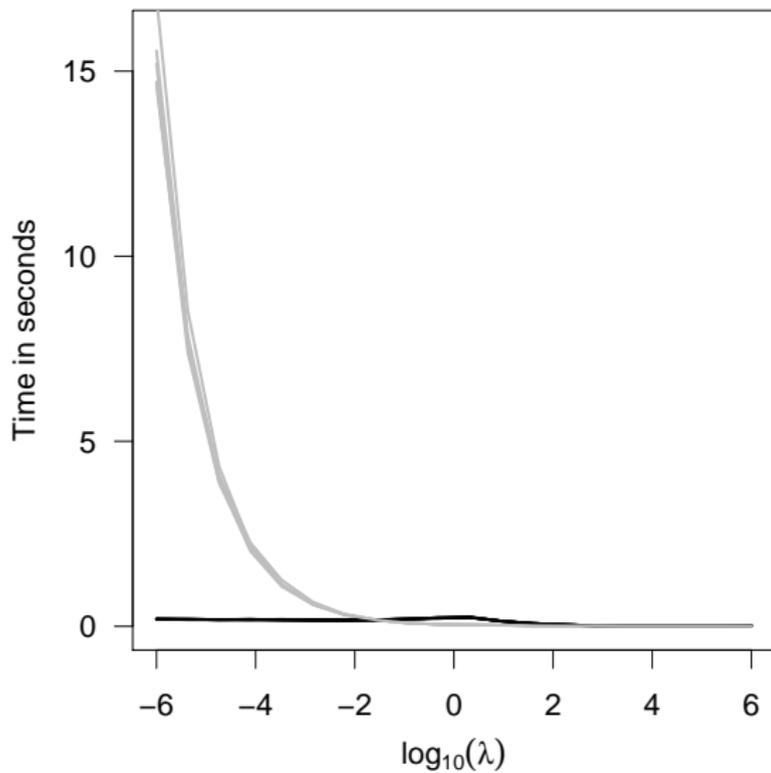
In each replication,

- ① randomly generated Σ_* :
 - eigenvectors were the right singular vectors of $Z \in \mathbb{R}^{p \times p}$, where the z_{ij} were drawn independently from $N(0, 1)$;
 - eigenvalues drawn independently from the uniform distribution on $(1, 100)$.
- ② generated S from an iid sample of size $n = 50$ from $N_p(0, \Sigma_*)$

Compared our MM algorithm to SPICE when computing $\hat{\Omega}_{\lambda, 2}(M_{\text{off}})$.

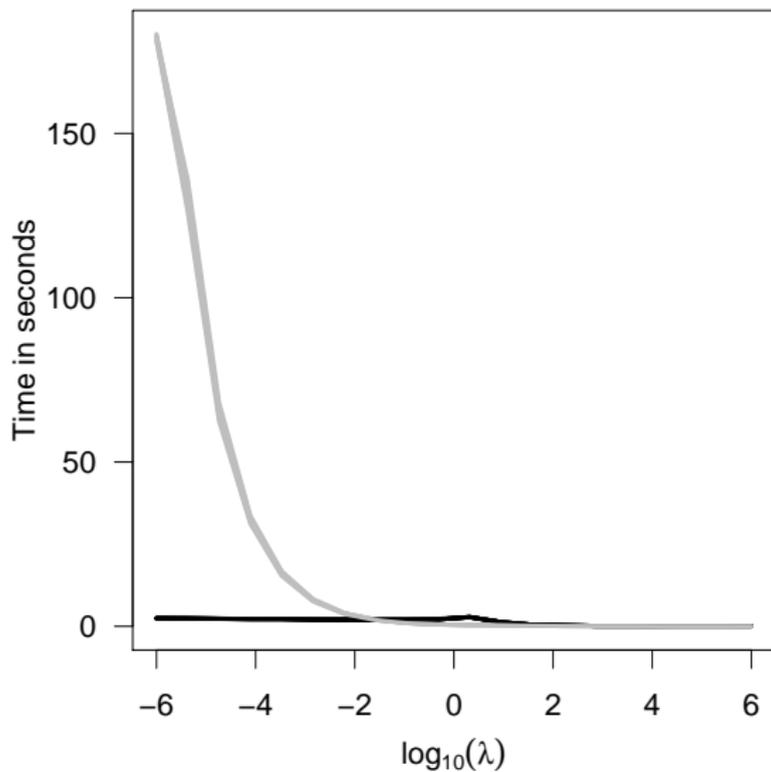
Results for $n = 50$ & $p = 100$

$$\hat{\Omega}_{\lambda,2}(M_{\text{off}})$$



Results for $n = 50$ & $p = 200$

$$\hat{\Omega}_{\lambda,2}(M_{\text{off}})$$



Tuning parameter selection

J -fold cross validation maximizing validation likelihood

- Randomly partition the n observations into J subsets of equal size.
- Compute

$$\hat{\lambda} = \arg \min_{\lambda \in \mathcal{L}} \sum_{j=1}^J \left\{ \text{tr} \left(\hat{\Omega}_{\lambda}^{(-j)} S^{(j)} \right) - \log \det \left(\hat{\Omega}_{\lambda}^{(-j)} \right) \right\},$$

- $S^{(j)}$ is computed from observations inside the j th subset (centered by training sample means)
- $\hat{\Omega}_{\lambda}^{(-j)}$ is the inverse covariance estimate computed from the observations outside the j th subset
- \mathcal{L} is some user defined finite subset of \mathbb{R}_+ , e.g. $\{10^{-8}, 10^{-7.5}, \dots, 10^8\}$.

Classification example

Task: discriminate between metal and rocks using sonar data.

- The data were taken from the UCI machine learning data repository.
- Sonar was used to produce energy measurements at $p = 60$ frequency bands for rock and metal cylinder examples.
- There were 111 metal cases and 97 rock cases.
- Quadratic discriminant analysis, with regularized covariance estimators was applied.
- Performed leave-one-out cross validation to compare classification performance.

Results: QDA on the sonar data

The number of classification errors made in leave-one-out cross validation on 208 examples.

	M_{off}	M_{all}
L_1 standardized	33	34
L_1	33	35
ridge standardized	33	37
ridge	31	42

These methods outperform using S^{-1} , which had 50 errors, and using $(S \circ I)^{-1}$, which had 68 errors.

Thank you

This talk is based on:

Rothman, A. J. and Forzani, L. (2014) Properties of optimizations used in penalized Gaussian likelihood inverse covariance matrix estimation. *Submitted*

An R package will be available soon