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

Adam J. Rothman
School of Statistics
University of Minnesota

April 21, 2014, joint work with Liliana Forzani

Estimating large covariance matrices and their inverses

- The covariance matrix is a fundamental quantity in multivariate analysis.
- In general, the sample covariance matrix S performs poorly in high dimensions ($p \geq n$).
- Shrinkage or regularized estimators are used instead.

Estimating large covariance matrices and their inverses

Desirable properties of the estimator

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

Two different problems:

A. Estimating Σ

B. Estimating Σ^{-1}

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).
- 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).
- 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.

Exploiting variable ordering

when estimating the covariance matrix Σ

- Banding or tapering S (Furrer & Bengtsson, 2007; Bickel & Levina, 2008; Cai, Zhang, and Zhou, 2010).
- Block thresholding (Cai & Yuan, 2012).
- Regularized estimation of the modified Cholesky factor of the covariance matrix (Rothman, Levina, & Zhu, 2010).

when estimating the inverse covariance matrix Σ^{-1}

- Regularized estimation of the modified Cholesky factor of the inverse covariance (Wu & Pourahmadi, 2003; Smith & Kohn, 2002; Bickel & Levina, 2008; Huang et al., 2006; Levina, Rothman, & Zhu, 2008).

Estimating Σ^{-1} in applications

- Classification (Bickel & Levina, 2004)
- 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

- $(X_1, \ldots, X_p)' \sim N_p(0, \Sigma), \quad \Sigma \in \mathbb{S}_+^p$.
- The undirected graph $G = (V, E)$ has
  - vertex set $V = \{1, \ldots, p\}$ 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

$$\hat{\Omega} = \arg\min_{\Omega \in \mathbb{S}^p_+} \{\mathrm{tr}(\Omega S) - \log|\Omega|\}$$

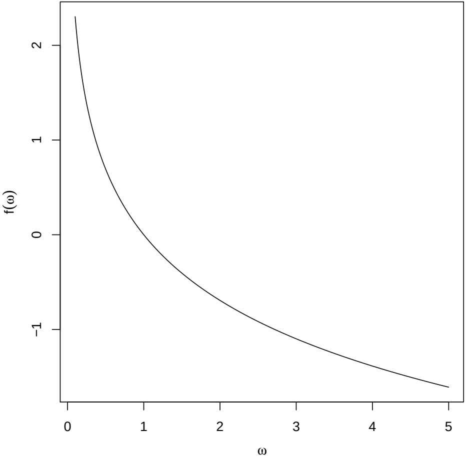
Set the gradient at $\hat{\Omega}$ to zero:

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

- 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 $f(\omega) = 0\omega - \log \omega$.



Penalized Gaussian likelihood estimation of Σ^{-1}

We will study

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

- $q = 1$ is the *lasso penalty*, $q = 2$ is the *ridge penalty*
- M is user-specified, symmetric with non-negative entries.
- The m_{ij}'s allow the user to incorporate prior information, e.g. it is known that $(\Sigma^{-1})_{21} = (\Sigma^{-1})_{12} \neq 0$ so set $m_{21} = m_{12} = 0$.
- Other examples M_{all}, M_{off}, and $m_{ij} = 1(|i-j| > k)$.
- We study when solutions exist and develop a new algorithm for the ridge penalty ($q = 2$).

Lasso penalized likelihood estimation of Σ^{-1}

The case when $q = 1$ is

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

- $\hat{\Omega}_{\lambda,1}(M_{\text{off}})$ Yuan & Lin (2007), Rothman et al. (2008), Lam & Fan (2009), and Ravikumar et al. (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 $\hat{\Omega}_{\lambda,1}(M_{\text{all}})$ exists if $\lambda > 0$.
- Ravikumar et al. (2008) showed that $\hat{\Omega}_{\lambda,1}(M_{\text{off}})$ exists if $\lambda > 0$ and $S \circ I \in \mathbb{S}^p_+$.
- Lu (2008) showed that $\hat{\Omega}_{\lambda,1}(M)$ exists if $S + \lambda M \circ I \in \mathbb{S}^p_+$.

Lasso-penalized likelihood solution existence

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

Theorem. The solution $\hat{\Omega}_{\lambda,1}(M)$ exists if and only if $A_{1,\lambda}(M) = \{\Sigma \in \mathbb{S}_+^p : |\sigma_{ij} - s_{ij}| \le m_{ij}\lambda\}$ 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. $\mathrm{soft}(S, \lambda M) \in \mathbb{S}_+^p$.

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

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

Ridge penalized likelihood estimation of Σ^{-1}

$$\hat{\Omega}_{\lambda,2}(M) = \arg\min_{\Omega \in \mathbb{S}^p_+} \left\{ \mathrm{tr}(\Omega S) - \log|\Omega| + 0.5\lambda \sum_{i,j} m_{ij}\omega_{ij}^2 \right\}$$

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

Ridge-penalized likelihood solution existence

$$\hat{\Omega}_{\lambda,2}(M) = \arg\min_{\Omega \in \mathbb{S}_+^p} \left\{ \operatorname{tr}(\Omega S) - \log|\Omega| + 0.5\lambda \sum_{i,j} m_{ij}\omega_{ij}^2 \right\}$$

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

Theorem. Suppose $\lambda > 0$. The solution $\hat{\Omega}_{\lambda,2}(M)$ exists if and only if $A_2(M) = \{\Sigma \in \mathbb{S}_+^p : \sigma_{ij} = s_{ij} \text{ when } (i,j) \in \mathcal{U}\}$ is not empty.

Remark. The MLE of the Gaussian graphical model with edge set $\mathcal{U}$ exits if and only if $A_2(M)$ is not empty (Buhl, 1993; Uhler, 2012)

Chordal graph example

From the theorem $\hat{\Omega}_{\lambda,2}(M)$ exists if and only if the MLE of the Gaussian graphical model with edge set $\mathcal{U} = \{(i,j) : m_{ij} = 0\}$ 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 $\hat{\Omega}_{\lambda,2}(M)$ exists if and only if $n \geq k+2$.

This follows from

Theorem[Buhl, 1993; Uhler, 2012]. Suppose that $\mathcal{U}$ is the edge set for a Chordal graph with maximum clique size q. The MLE of the zero mean Gaussian graphical model with edge set $\mathcal{U}$ exists if and only if $n \geq q$.

Entry agreement between $\hat{\Omega}_{\lambda,2}(M)^{-1}$ and S

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

Remark. When $\hat{\Omega}_{\lambda,2}(M)$ exists,

$$\{\hat{\Omega}_{\lambda,2}(M)^{-1}\}_{ij} = s_{ij} \quad \text{for } (i,j) \in \mathcal{U}.$$

Justification. The zero gradient equation is

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

Ridge and lasso connections

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

$$A_{1,\lambda}(M) = \{\Sigma \in \mathbb{S}_+^p : |\sigma_{ij} - s_{ij}| \leq m_{ij}\lambda\}$$
$$A_2(M) = \{\Sigma \in \mathbb{S}_+^p : \sigma_{ij} = s_{ij} \text{ when } (i,j) \in \mathcal{U}\}$$

- $A_{1,\lambda}(M) \subset A_2(M)$.
- If $\hat{\Omega}_{\lambda,1}(M)$ exists, then $\hat{\Omega}_{\tilde{\lambda},2}(M)$ exists for all $\tilde{\lambda} > 0$.
- If $\hat{\Omega}_{\lambda,2}(M)$ exists for some $\lambda > 0$, then it exists for all $\lambda > 0$, 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 Ridge solution

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

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

- Each iteration minimizes

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

  - 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,

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

- At the kth 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

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

  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).

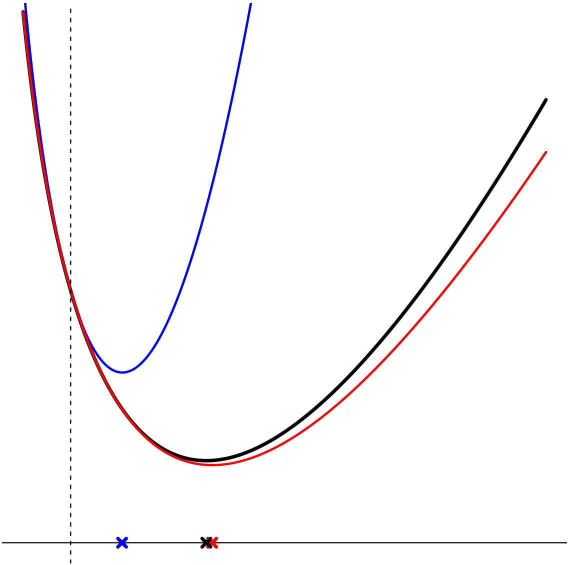
Acceleration by minorizing f at Ω_k

- 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 $\hat{\Omega}_{\lambda,2}(M)$ exits, then

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

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

Simulation: our algorithm vs the SPICE algorithm

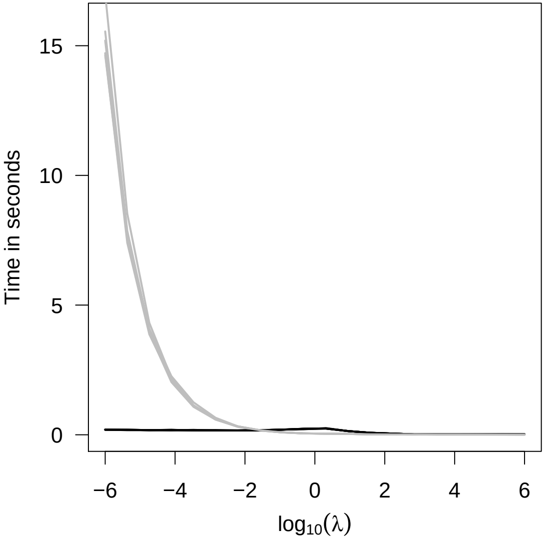
In each replication,

1. randomly generated Σ_0:
   - 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 $N_p(0, \Sigma_0)$

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

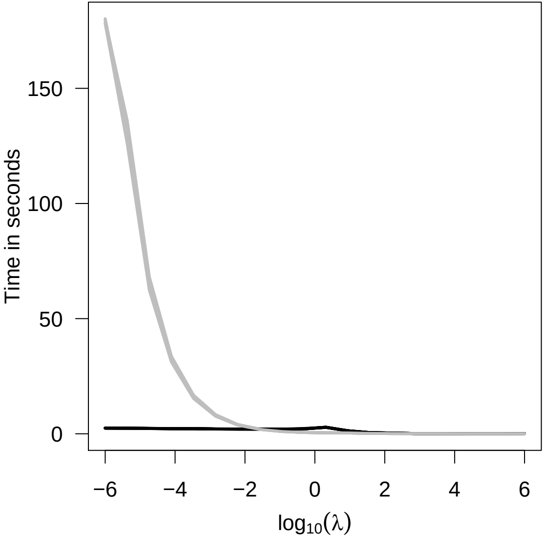
Results for $n = 50$ & $p = 100$

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



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

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

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

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. (2013) Properties of optimizations used in penalized Gaussian likelihood inverse covariance matrix estimation. *Submitted*

An R package will be available soon