Regularized estimation of large covariance matrices

Liza Levina
Department of Statistics
University of Michigan

Parts joint work with
Peter Bickel (UC Berkeley)
Ji Zhu (University of Michigan)
Adam Rothman (University of Michigan)

Why estimate covariance?

- Principal component analysis (PCA)
- Linear or quadratic discriminant analysis (LDA/QDA)
- Inferring independence and conditional independence in Gaussian graphical models
- Inference about the mean (e.g., longitudinal mean response curve)

Covariance itself is not always the end goal:
- PCA requires estimation of the eigenstructure
- LDA/QDA and conditional independence require the inverse

What’s wrong with the sample covariance?

Observe \( X_1, \ldots, X_n \), i.i.d. \( p \)-variate random variables

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T
\]

- MLE, unbiased (almost), well-behaved (and well studied) for fixed \( p \), \( n \to \infty \). But very noisy if \( p \) is large.
- Eigenvectors are not consistent (Johnstone and Lu, 2004; Paul, 2006)
- LDA breaks down if \( p/n \to \infty \) (Bickel and Levina, 2004)
- Singular if \( p > n \)

Desirable features for a covariance estimator

- Positive definite
- Well-conditioned
- Inverse easily available (in some applications)
- Sparse covariance or inverse (in some applications)
- Ultimately, consistency and good rates for large \( p \)
Two classes of estimators

- Variables have a natural notion of distance (time series, spatial data, longitudinal data, spectroscopy, ...)
  - exploit weaker dependence between “distant” variables
- The variable order is meaningless (genetics, social science, demographics, finance)
  - estimators should be invariant to variable permutations

Early alternatives: Steinian shrinking of eigenvalues

- First proposed by Stein (Rietz lecture, 1975)
- Empirical Bayes (Haff, 1980), minimax shrinkage (Dey and Srinivasan, 1985)
- Ledoit and Wolf (2003): $\rho_1 \hat{\Sigma} + \rho_2 I$, optimal $\rho_1, \rho_2$ estimated
- Shrinkage estimators solve the singularity problem but are not sparse and do not change eigenvectors.
- Bayesian shrinkage done via priors (Daniels and Kass, 2001; Smith and Kohn, 2002; ...)

Outline

- Introduction
- Regularizing covariance when variables are ordered
- Regularizing the inverse covariance when variables are ordered
- Inverse regularization invariant to variable permutations (SPICE)

Banding (tapering) of covariance matrices

- Replace $\hat{\Sigma}$ with $\hat{\Sigma} \ast R$, where $\ast$ means Schur (element-wise) product
- If $R$ is positive definite, so is $\hat{\Sigma} \ast R$

Examples:

- Banding (not positive definite):
  $$R_k(i, j) = 1(|i - j| \leq k)$$

- “Triangular” filter: banded, positive definite
  $$R_k(i, j) = \left(1 - \frac{|i - j|}{k + 1}\right)_+$$

- “Exponential” filter: positive definite but not banded
  $$R_\sigma(i, j) = e^{-\frac{|i - j|}{\sigma}} = \rho^{|i - j|}$$
Convergence of banded estimators

Bickel and Levina (2007)

- All results in operator norm, a.k.a. matrix 2-norm: for symmetric $M$,

$$\|M\| = \max_i |\lambda_i(M)|$$

- Result uniform over a class of covariance matrices as $p, n \to \infty$

$$U(\varepsilon_0, \alpha, C) = \{ \Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0, \max_j \sum_i |\sigma_{ij}| : |i - j| > k \leq Ck^{-\alpha} \text{ for all } k \geq 0 \}.$$  

- The class includes stationary processes with bounded smooth spectral density + well-behaved non-stationary noise.

Extensions

- Gaussianity may be replaced by a tail condition

- The theorem also holds for $\hat{\Sigma} * R_{\sigma}$, where

$$R_{\sigma}(i, j) = g \left( \frac{\rho(i, j)}{\sigma} \right)$$

where $\rho$ is a metric on the set of variable labels, $g$ is a tapering function (continuous, non-increasing, $g(0) = 1$, $g(\infty) = 0$), and $\sigma > 0$.

- includes triangular and exponential filters

- Convergence in operator norm implies convergence of eigenvalues and eigenvectors (el Karoui, 2007).

Banded estimator:

$$\hat{\Sigma}_{k,p}(i, j) = \hat{\Sigma}_p(i, j) \cdot 1(|i - j| \leq k)$$

Theorem: If $X$ is Gaussian and $k \asymp (n^{-1} \log p)^{-1/(\alpha+1)}$, then, uniformly on $\Sigma \in U(\varepsilon_0, \alpha, C)$,

$$\|\hat{\Sigma}_{k,p} - \Sigma_p\| = O_P \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha+1)}} = \|\hat{\Sigma}_{k,p}^{-1} - \Sigma_p^{-1}\|$$

The banded estimator and its inverse are consistent if $\frac{\log p}{n} \to 0$. 

- Implicitly assume $|i - j|$ large implies $X_i$ and $X_j$ nearly uncorrelated

- Bickel and Levina (2004): banding Toeplitz matrices leads to convergence to Bayes risk for LDA
  - not evaluated in the context of general estimation

- Furrer and Bengtsson (2006): tapering covariance in the context of Kalman filtering
  - $R$ is a function of $\Sigma$

- All results in operator norm, a.k.a. matrix 2-norm: for symmetric $M$,

$$\|M\| = \max_i |\lambda_i(M)|$$

- Result uniform over a class of covariance matrices as $p, n \to \infty$

$$U(\varepsilon_0, \alpha, C) = \{ \Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0, \max_j \sum_i |\sigma_{ij}| : |i - j| > k \leq Ck^{-\alpha} \text{ for all } k \geq 0 \}.$$  

- The class includes stationary processes with bounded smooth spectral density + well-behaved non-stationary noise.

Banded estimator:

$$\hat{\Sigma}_{k,p}(i, j) = \hat{\Sigma}_p(i, j) \cdot 1(|i - j| \leq k)$$

Theorem: If $X$ is Gaussian and $k \asymp (n^{-1} \log p)^{-1/(\alpha+1)}$, then, uniformly on $\Sigma \in U(\varepsilon_0, \alpha, C)$,

$$\|\hat{\Sigma}_{k,p} - \Sigma_p\| = O_P \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha+1)}} = \|\hat{\Sigma}_{k,p}^{-1} - \Sigma_p^{-1}\|$$

The banded estimator and its inverse are consistent if $\frac{\log p}{n} \to 0$. 

- Implicitly assume $|i - j|$ large implies $X_i$ and $X_j$ nearly uncorrelated

- Bickel and Levina (2004): banding Toeplitz matrices leads to convergence to Bayes risk for LDA
  - not evaluated in the context of general estimation

- Furrer and Bengtsson (2006): tapering covariance in the context of Kalman filtering
  - $R$ is a function of $\Sigma$
Choosing the tuning parameter

Want to minimize risk

\[ R(k) = E\|\hat{\Sigma}_k - \Sigma\| \]

Estimate via a resampling scheme (cross-validation):

- Split the data into two samples \( N \) times at random
- Let \( \hat{\Sigma}_1^{(\nu)}, \hat{\Sigma}_2^{(\nu)} \) be the two sample covariance matrices from the \( \nu \)-th split. The risk can be estimated by

\[
\hat{R}(k) = \frac{1}{N} \sum_{\nu=1}^{N} \| (\hat{\Sigma}_1^{(\nu)})_k - (\hat{\Sigma}_2^{(\nu)})_k \|
\]

Banding performance in simulations

- Picking \( k \) by cross-validation works well
- The same model requires more regularization in higher dimensions
- If conditions are not satisfied (e.g., long-range dependence), tend to pick \( k \approx p \)
- Useful for PCA as well as for general estimation

Outline

✔ Introduction
✔ Regularizing covariance when variables are ordered (banding)
- Regularizing the inverse covariance when variables are ordered (banding, adaptive banding)
- Inverse regularization invariant to variable permutations (SPICE)

Cholesky decomposition of covariance

- Any \( p \)-variate \( \mathbf{X} \) with mean 0 and covariance \( \Sigma \): regress \( X_j \) on \( X_{j-1}, \ldots, X_1 \)

\[
\hat{X}_j = \sum_{t=1}^{j-1} \phi_{jt} X_t, \quad \varepsilon_j = X_j - \hat{X}_j, \quad d_j^2 = \text{Var}(\varepsilon_j)
\]

- Let \( \mathbf{A} = [\phi_{jt}] \) (lower triangular), \( \mathbf{T} = \mathbf{I} - \mathbf{A} \), \( \mathbf{D} = \text{diag}(d_j^2) \). Write

\[
\varepsilon = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{T} \mathbf{X}
\]

Independence of residuals \( \Rightarrow \) modified Cholesky decomposition:

\[
\mathbf{D} = \mathbf{T} \Sigma \mathbf{T}^T, \quad \Sigma^{-1} = \mathbf{T}^T \mathbf{D}^{-1} \mathbf{T}
\]

- Transforms covariance estimation into a regression problem
Likelihood as a function of Cholesky parameters

Assuming normality, the negative log-likelihood can be written as

\[
\ell(X, \Sigma) = n \ln |\Sigma| + \sum_{i=1}^{n} x_i^T \Sigma^{-1} x_i = \sum_{j=1}^{p} \ell_j(x_1, \ldots, x_n, T, D)
\]

where

\[
\ell_1(X, T, D) = \ell_1(X, d_1) = \ln d_1^2 + \frac{1}{d_1^2} \sum_{i=1}^{n} x_{i1}^2
\]

\[
\ell_j(X, T, D) = \ell_j(X, \phi_j, d_j) = \ln d_j^2 + \frac{1}{d_j^2} \sum_{i=1}^{n} (x_{ij} - \sum_{t=1}^{j-1} \phi_{jt} x_{it})^2
\]

Each \( \ell_j \) can be minimized separately.

Estimators based on Cholesky decomposition

- Shrink elements of \( T \) using regression methods (otherwise just get sample \( \hat{\Sigma} \)).
- Are always positive definite \( (\Sigma^{-1} = T^T D^{-1} T) \).
- Not invariant under variable permutations.
- Give a natural estimate of \( \Sigma^{-1} \) rather than \( \Sigma \).
- Wu and Pourahmadi (2003): Banding the Cholesky factor – regress on \( k \) predecessors, smooth sub-diagonals with a spline.
- Huang, Liu, Pourahmadi, and Liu (2006): fit \( T \) and \( D \) by maximum likelihood penalized by lasso or ridge penalty on \( \phi_{jt} \).

Banding the Cholesky factor

- Center variables
- Regress \( X_j \) on \( X_{j-1}, \ldots, X_{j-k} \); get new matrices of coefficients \( \hat{A}_k \), and residual variances \( \hat{D}_k \)
- Define \( \hat{T}_k = I - \hat{A}_k \), and let

\[
\hat{\Sigma}_k^{-1} = \hat{T}_k^T \hat{D}_k^{-1} \hat{T}_k,
\]

\[
\hat{\Sigma}_k = [\hat{\Sigma}_k^{-1}]^{-1}.
\]

- \( \hat{\Sigma}_k^{-1} \) is \( k \)-banded nonnegative definite; \( \hat{\Sigma}_k \) is in general not banded.

Convergence of the banded Cholesky estimator

Bickel and Levina (2007)

Define a class of covariance matrices: if \( \Sigma^{-1} = T(\Sigma)^T D(\Sigma)^{-1} T(\Sigma) \),

\[
U^{-1}(\varepsilon_0, C, \alpha) = \{ \Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \varepsilon_0, \max_i \sum_{j<i-k} |t_{ij}(\Sigma)| \leq C k^{-\alpha} \text{ for all } k \leq p-1 \}
\]

Theorem: Uniformly for \( \Sigma \in U^{-1}(\varepsilon_0, C, \alpha) \), if \( X \) is Gaussian, \( k \asymp \left( \frac{\log p}{n} \right)^{-\frac{1}{2(\alpha+1)}} \), and \( \frac{\log p}{n} = o_P(1) \),

\[
\|\hat{\Sigma}_{k,p}^{-1} - \Sigma_p^{-1}\| = O_P \left( \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha+1)}} \right) = \|\hat{\Sigma}_{k,p} - \Sigma_p\|.
\]
Adaptive banding of the Cholesky factor

Levina and Zhu (2006)

Recall each $\ell_j$ can be minimized separately. To force shrinkage, minimize

$$\min_{\phi_j, d_j} \ell_j(X, \Sigma) + J(\phi_j)$$

LASSO penalty (Huang et al., 2006):

$$J(\phi_j) = \lambda \sum_{t=1}^{j-1} |\phi_{jt}|$$

- Shrinkage + sparsity: some $\hat{\phi}_{jt} = 0$
- Sparse in $T$, not necessarily in $\Sigma^{-1} = T^TD^{-1}T$

Nested LASSO penalty

$$J(\phi_j) = \lambda_1 \sum_{t=1}^{j-1} |\phi_{jt}| + \lambda_2 \sum_{t=1}^{j-2} \frac{|\phi_{jt}|}{|\phi_{jt+1}|}$$

- Shrinkage + sparsity: If $\hat{\phi}_{jt} = 0 \implies \hat{\phi}_{jj'} = 0$ for all $j' < t$.
- Sparse in $T$ and $\Sigma^{-1}$
- Nested LASSO $\implies$ Adaptive Banding of $\Sigma^{-1}$

Simulation results

- Two models
  $\Sigma_1$: $\phi_{j,j-1} = 0.8$ and $\phi_{jj'} = 0$ otherwise (tri-diagonal inverse);
  $$d_j = 0.1.$$  
  $\Sigma_2$: $k_j \sim U(1, [j/2])$; $\phi_{jj'} = 0.5$, $j' \geq k_j$ and $\phi_{jj'}$ otherwise;
  $$d_j = 0.1.$$  
- Kullback-Leibler loss (Yuan and Lin, 2007)
  $$\Delta_{KL}(\Sigma, \hat{\Sigma}) = \text{tr} \left( \hat{\Sigma}^{-1} \Sigma \right) - \ln \left| \hat{\Sigma}^{-1} \Sigma \right| - p$$
- $n = 100$, validation set size 100 (for selecting $\lambda$), loss averaged over 50 replications
**Preserving sparsity in the Cholesky factor**

Percentage of zeros out of 50 replications

<table>
<thead>
<tr>
<th></th>
<th>Sample</th>
<th>Ledoit-Wolf</th>
<th>Lasso</th>
<th>Adapt. Banding</th>
<th>Banding</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>8.38(0.97)</td>
<td>3.59(0.30)</td>
<td>1.26(0.27)</td>
<td>0.64(0.12)</td>
<td>0.63(0.13)</td>
</tr>
<tr>
<td>100</td>
<td>NA</td>
<td>29.33(0.88)</td>
<td>6.91(0.79)</td>
<td>2.21(0.23)</td>
<td>2.21(0.23)</td>
</tr>
<tr>
<td>200</td>
<td>NA</td>
<td>90.86(1.32)</td>
<td>14.57(0.89)</td>
<td>4.35(0.37)</td>
<td>4.34(0.38)</td>
</tr>
<tr>
<td>$\Sigma_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>8.68(0.87)</td>
<td>171.31(7.06)</td>
<td>4.62(0.52)</td>
<td>3.14(0.41)</td>
<td>3.82(0.34)</td>
</tr>
<tr>
<td>100</td>
<td>NA</td>
<td>945.65(15.20)</td>
<td>35.60(5.05)</td>
<td>11.24(0.87)</td>
<td>14.34(0.64)</td>
</tr>
<tr>
<td>200</td>
<td>NA</td>
<td>1938.32(21.47)</td>
<td>118.84(10.86)</td>
<td>22.70(1.15)</td>
<td>29.50(1.02)</td>
</tr>
</tbody>
</table>

**Preserving sparsity in the inverse**

**Classification example: Mass spectroscopy data**

- 157 healthy blood samples, 167 with prostate cancer
- Intensities at close mass/charge ratios are strongly correlated
- 48538 original sites reduced to 218 by block averaging
- Compare test error for LDA for 6 covariance estimators

<table>
<thead>
<tr>
<th></th>
<th>Sample</th>
<th>Naive Bayes</th>
<th>Ledoit &amp; Wolf</th>
<th>Banding</th>
<th>Lasso</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.13</td>
<td>0.30</td>
<td>0.15</td>
<td>0.19</td>
<td>0.18</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Outline

✔ Introduction

✔ Regularizing covariance when variables are ordered (banding)

✔ Regularizing the inverse covariance when variables are ordered (banding, adaptive banding)

● Inverse regularization invariant to variable permutations (SPICE)

Estimators of the inverse invariant under variable permutations

● Inverse $\Omega = \Sigma^{-1}$ (a.k.a. concentration matrix or precision matrix)

● Graphical models focus on testing/finding zeros rather than estimating the matrix (Drton and Perlman 2007; Meinshausen and Bühlmann 2006)

● Also needed for LDA/QDA

Sparse permutation-invariant covariance estimator

● The negative log-likelihood up to a constant:

$$\ell(X; \Omega) = \text{tr}(\Omega \hat{\Sigma}) - \log |\Omega|$$

● Add lasso penalty on off-diagonal elements of $\Omega \Rightarrow$ force regularization and sparsity.

$\textbf{SPICE:}$  $\hat{\Omega}_\lambda = \arg \min_{\Omega > 0} \{ \text{tr}(\Omega \hat{\Sigma}) - \log |\Omega| + \lambda \sum_{j \neq j'} |\Omega_{jj'}| \}$

● Optimization is non-trivial and computationally expensive

  – Yuan and Lin (2007): Maxdet algorithm (interior point optimization), a convergence result for fixed $p$

  – Banerjee et al. (2006): Nesterov’s convex optimization to solve a similar problem (penalty includes diagonal elements)

Analysis of SPICE

Rothman, Bickel, Levina, Zhu (2007)

Theorem: Let the true concentration matrix be $\Omega_0 = \Sigma_0^{-1}$. Assume

A1: $0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma_0) \leq \lambda_{\max}(\Sigma_0) \leq \varepsilon_0^{-1}$

A2: $\text{card}\{ (i, j) : \Omega_{0ij} \neq 0, i \neq j \} \leq s$

Then if $\lambda \asymp \frac{\sqrt{\log p}}{n}$,

$$\| \hat{\Omega}_\lambda - \Omega_0 \|_F = O_P \left( \sqrt{\frac{(p + s) \log p}{n}} \right).$$

● Uses Frobenius norm $\| M \|_F^2 = \sum_{i,j} m_{ij}^2$.

● $\sqrt{\frac{(1+s) \log p}{n}}$ in operator norm working with the correlation matrix

● Still better than the sample covariance rate of $\frac{p}{\sqrt{n}}$. 
**Cholesky-based optimization algorithm**

- Want to avoid semi-definite programming
- Idea: re-parametrize the objective using the Cholesky decomposition
  \[ \Omega = T^T T, \]
  where \( T = [t_{ij}] \) is a lower triangular matrix.
- Positive definiteness is automatic

**Objective becomes**

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( \sum_{k=1}^{j} t_{jk} X_{ik} \right)^2 - 2 \sum_{j=1}^{p} \log t_{jj} + 2\lambda \sum_{j'>j}^{p} \sum_{k=j'}^{p} t_{kk} t_{k'j'}
\]

- Update one column of \( T \) at a time ⇒ product terms separate
- Again replace absolute value by quadratic approximation
- Now both likelihood and penalty are quadratic in \( t_{ij} \) ⇒ take derivatives and solve a linear system in \( t \)
- Sweep through columns and iterate till convergence
- Not convex, but at each step the objective function decreases

**SPICE simulation example**

Two models:

1. \( \Omega_1: \phi_{j1} = 0.8; \phi_{jj'} = 0, j' > 1; d_j = 0.1. \)
   This corresponds to \( X_1 = \varepsilon_1, X_j = 0.8X_1 + d_j \varepsilon_j \) for \( j = 2, \ldots, p, \) with \( \varepsilon_j \) i.i.d. \( N(0,1). \)

2. \( \Omega_2 \) corresponds to a permutation of the same process:
   \( X_2, \ldots, X_p, X_1. \) Now \( \phi_{jj'} \neq 0 \) for all \( j, j'. \)
- Both \( \Omega_1 \) and \( \Omega_2 \) are sparse, but only \( \Omega_1 \) has a sparse Cholesky factor.

- \( n = 100, \) validation set size 100 (for selecting \( \lambda \)), loss averaged over 50 replications
- Kullback-Leibler loss

<table>
<thead>
<tr>
<th>( p )</th>
<th>Sample</th>
<th>Ledoit-Wolf</th>
<th>SPICE</th>
<th>Lasso ( \Omega_1 )</th>
<th>Lasso ( \Omega_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>8.38(0.97)</td>
<td>5.34(0.47)</td>
<td>1.92(0.20)</td>
<td>2.99(0.35)</td>
<td>3.80(0.48)</td>
</tr>
<tr>
<td>100</td>
<td>NA</td>
<td>116.64(13.02)</td>
<td>7.02(0.55)</td>
<td>19.13(2.20)</td>
<td>25.56(2.63)</td>
</tr>
</tbody>
</table>
Heatmaps of zeros identified in $\Omega_1$ and $\Omega_2$

(a) True $\Omega_1$  
(b) Lasso $\hat{\Omega}_1$  
(c) SPICE $\hat{\Omega}_1$

(d) True $\Omega_2$  
(e) Lasso $\hat{\Omega}_2$  
(f) SPICE $\hat{\Omega}_2$

Colon tumor classification example

- 40 colon adenocarcinoma and 22 healthy tissue samples (Alon et al, 1999)
- Affymetrix gene expression data
- Top $p$ genes selected by two-sample $t$-test (out of 2000)
- 100 random splits into training (2/3) and test (1/3) sets
- LDA classification
- Consider two ways of selecting the tuning parameter

Colon data classification errors

Tuning parameter for Lasso and SPICE chosen by

(A) 5-fold CV on the training data maximizing the likelihood;

(B) 5-fold CV on the training data minimizing the classification error;

(C) minimizing the classification error on the test data

<table>
<thead>
<tr>
<th>$p$</th>
<th>N. Bayes</th>
<th>L-W</th>
<th>Lasso A</th>
<th>Lasso B</th>
<th>Lasso C</th>
<th>SPICE A</th>
<th>SPICE B</th>
<th>SPICE C</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>15.8(7.7) 15.2(5.5)</td>
<td>15.3(7.2) 34.3(19.7) 12.0(6.4)</td>
<td>12.1(6.5)</td>
<td>14.7(7.3)</td>
<td>9.0(5.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20.0(8.4) 16.3(7.1)</td>
<td>19.5(8.2) 38.2(14.2) 16.4(7.3)</td>
<td>18.7(8.4)</td>
<td>16.9(8.5)</td>
<td>9.1(5.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>24.4(9.7) 18.1(6.5)</td>
<td>24.3(9.9) 41.0(14.9) 18.7(7.2)</td>
<td>18.9(6.6)</td>
<td>18.6(7.3)</td>
<td>10.7(6.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

- Regularization is necessary
- High-dimensional asymptotics give rates of $p$ vs. $n$ and conditions on the covariance model
- Penalties can be used to impose desired structure
- Efficient optimization is very important
- Cholesky decomposition is a useful tool even with no ordering