Regularized Estimation of Covariance Matrices and Its Uses

Peter Bickel

NCAR

May, 2007

Outline

1. Uses of covariance matrix estimates in climate research
2. Pathologies of high dimensional empirical covariance matrices
3. Sparsity
   a) in EOF's
   b) in ensemble covariance matrices
4. Regularization by "banding" of entries of empirical matrix
5. Choice of band size
6. Some simulation
7. PCA(EOF)

Covariance Matrices in Climate Research I

Zwiers and Von Storch (1999), Statistical analysis in climate research.

- Empirical Orthogonal Functions(EOF) aka principal components.
- $\mathbf{X}_1, \ldots, \mathbf{X}_n$: p dimensional stationary vector time series.
- *e.g.* Zwiers and Von Storch (1999) Chapter 13. sea surface monthly
  - Average anomaly (107 years, $5^o \times 5^o$ grid, 360^o latitude, 100^o longitude)
  - $p = 72 \times 20 = 1440$, $n = 107$

Covariance Matrices in Climate Research II

- $\mathbf{X}_k = \{X(i,j) \ : \ i = \text{latitude}, \ j = \text{longitude}\}$
- $\mathbb{X}_{n\times p} = \left(\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T\right)^T$
- $\mathbb{E}(\mathbf{X}) = \mathbf{0}$
- $\Sigma = Var(\mathbf{X}) = \mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right) = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j^T$
- $\mathbf{e}_1, \ldots, \mathbf{e}_p$: Principal components.
- $\lambda_1 > \cdots > \lambda_p$: Eigenvalues.
- Goal : Estimate, interpret $\mathbf{e}_j$, $j = 1, \ldots, K$

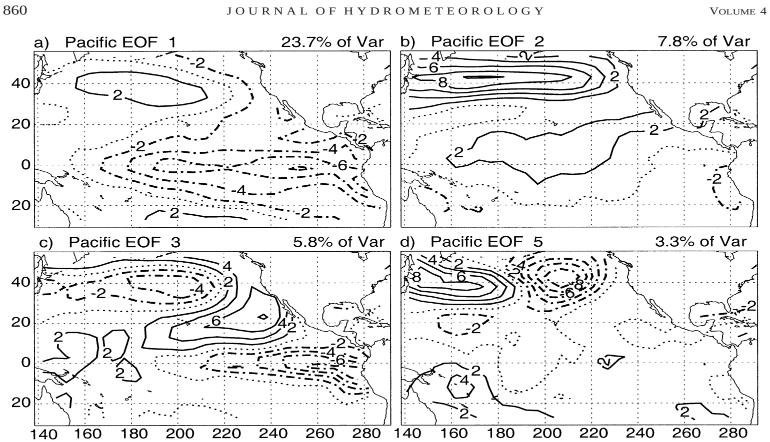
  such that $\dfrac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$ large.

Covariance Matrices in Climate Research III

Sparsity

G. R. Markowski and G. R. North

Journal of Hydrometeorology (2003).



Example : National Centre for Atmospheric Research I

- Computer model
  - X_j = ave. "pressure", "temperature",... in $50km \times 50km \times$ *variable* block of atmosphere, $|J| \asymp 10^7$, computer model.
  - $\mathbf{X}_i = \mathbf{X}(t_i)$ $i = 1, \ldots, T$.
  - Theory and practice suggest that $X_j(t_i)$ and $X_k(t_i)$ essentially independent if blocks j and k are far from each other.

Example : National Centre for Atmospheric Research II

- Data assimilation
  - $\mathbf{Y}(t_i)$: Data vectors.
  - Ensemble: $\mathbf{X}_j^U(t)$, $1 \leq j \leq n$.
  - Data assimilated : $\mathbf{X}_j^F(t)$, $1 \leq j \leq n$ uses $\hat{\Sigma}^{-1}$ as estimate of true $\left[Var(\mathbf{X}_1^U)\right]^{-1}$ for Kalman gain.

Pathologies of empirical covariance matrix

Observe $\mathbf{X}_1, \ldots, \mathbf{X}_n$, i.i.d. p-variate random variables

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T$$

- MLE, for Gaussian unbiased (almost), well-behaved (and well studied) for fixed p, $n \to \infty$. But very noisy if p is large.
- Singular if $p > n$, so $\hat{\Sigma}^{-1}$ is not uniquely defined.
- Computational issues with $\hat{\Sigma}^{-1}$ for large p.
- LDA completely breaks down if $p/n \to \infty$
- Eigenstructure inconsistent as soon as $p/n \to c > 0$.

Eigenvalues

Description of spreading phenomenon:

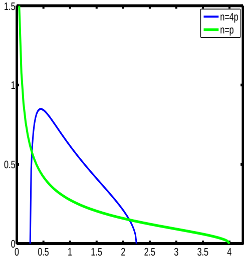
Empirical distribution function: for eigenvalues $\{\hat{\ell}_i\}_{j=1}^{p}$

$$G_p(t) = p^{-1}\#\{\hat{\ell}_j \leq t\} \rightarrow G(t) \leftrightarrow g(t)dt.$$

Marčenko-Pastur, (67), For $A \sim W_p(n, I)$ $\quad p/n \rightarrow \gamma$

For $\Sigma = I$,

$$g^{MP}(t) = \frac{\sqrt{(b_+ - t)(t - b_-)}}{2\pi\gamma t},$$
$$b_\pm = (1 \pm \sqrt{\gamma})^2.$$



Eigenvectors

D.Paul (2006) For the spike model

$$\Sigma = diag(\lambda, 1, \ldots, 1) = \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & & \ddots & \\ 0 & & \cdots & 1 \end{pmatrix}, 1 < \lambda < 1 + \sqrt{\gamma},$$

Eigenvector $\hat{\mathbf{e}} \leftrightarrow \hat{\lambda}$, $\mathbf{e} \leftrightarrow \lambda$,

$$|\mathbf{e} - \hat{\mathbf{e}}|^2 \xrightarrow{a.s.} 2$$

An Orthogonal Factor Model

- Model : $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. $N(\mu, \Sigma)$.

$$\Sigma = \Sigma_0 + \sigma^2 I_p, \quad \text{where} \quad \Sigma_0 = \sum_{j=1}^{M} \lambda_j \boldsymbol{\theta}_j \boldsymbol{\theta}_j^t$$

and $\lambda_1 \geq \ldots \geq \lambda_M > 0$, $\{\boldsymbol{\theta}_j\}$ orthonormal.

- Equivalent :

$$\mathbf{X}_i = \mu + \sum_{j=1}^{M} \sqrt{\lambda_j} v_{ji} \boldsymbol{\theta}_j + \sigma \mathbf{Z}_i, \quad i = 1, \ldots, n, \ v_{ji} \ \sim \ i.i.d.\ \mathcal{N}(0,1)$$

Sparsity in EOF

- p, n large but,

  (i) M small fixed.

  (ii) $\boldsymbol{\theta}_j$ "sparse"

- "well approximated" by $\boldsymbol{\theta}_{js}$, $||\boldsymbol{\theta}_{js}||_0 \leq s$, s "small".

  $||\mathbf{v}_{js}||_0 \equiv \#$ of nonzero coordinates of $\mathbf{v}$.

▸ Figure

Sparsity II

- Label dependent not directly related to eigenstructure of covariance matrix.
- If $\Sigma = ||\sigma_{ij}||$,
  - $|\sigma_{ij}|$ small (effectively 0) for $|i - j|$ large
  - More generally, given metric m on J, $|\sigma_{ij}|$ small if $m(i,j)$ large
  - $\sigma_{ij} = 0 \Rightarrow X_i \perp X_j$ under Gaussianity.

Examples

(i) $\mathbf{X}$ stationary $\sigma(i,j) = \sigma(|i-j|)$

$\leftrightarrow$ spectral density f, $0 < \varepsilon \leq f \leq \frac{1}{\varepsilon} < \infty$

- Ergodic ARMA processes satisfy

(ii) $T = S + K \leftrightarrow \mathbf{X} = \mathbf{Y} + \mathbf{Z}$

$\mathbf{Y}, \mathbf{Z}$ independent, $S \leftrightarrow \mathbf{Y}$, $K \leftrightarrow \mathbf{Z}$, $S = [s(i-j)]$ as in (i),

K Hilbert-Schmidt:

$$\sum_{i,j} K^2(i,j) < \infty \ (Z_m \xrightarrow{p} 0, \text{ non stationary})$$

$$(i) \Rightarrow \sum_i s^2(i) < \infty$$

Sparsity III of inverse covariance matrix

If $\Sigma^{-1} = ||\sigma^{ij}||$,

- $|\sigma^{ij}| = 0$ for "many" (i, j) pairs if $|i - j|$ is large.

Implication:

$\sigma^{ij} = 0 \Rightarrow X_i \perp X_j \mid \{X_k : k \neq i, j\}$ for Gaussian case.

Sparsity IV

- Permutation invariant sparsities

  $a)$ Each row of Σ sparse or sparsely approximable

  e.g. If $\boldsymbol{\sigma}_i = \{\sigma_{i,j} : 1 \le j \le p\}$, $||\boldsymbol{\sigma}_i||_0 \le s$

  $b)$ Each row of Σ^{-1} sparsely approximable.

- *a*) roughly implies *b*) if $\lambda_{min}(\Sigma) \ge \delta > 0$.

- The graph with edge weight between i and j given by σ_{ij} (or σ^{ij}) is "sparse" in some suitable sense (El Karoui (2007)).

Graphical models

Meinshausen and Buhlmann (2006), Zhao and Yu (2006),

Wainwright (2006), Kalisch and Buhlmann (2007)

Σ^{-1} corresponds to a graphical model.

$\mathcal{N}(i) = \{j : \sigma^{ij} \neq 0, j \neq i\}$

Goal : Determine $\mathcal{N}(i)$ $i = 1, \ldots, p$.

Example : Gene networks.

Regularization of $\hat{\Sigma}$ by banding or tapering I

Bickel and Levina (2004,2006), Furrer and Bergtsson (2006)

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

Examples:

- Banding(not positive definite):

$$R_k(i,j) = \mathbf{1}(|i-j| \leq k)$$

- "Triangular" filter: banded, positive definite

$$R_k(i,j) = \left(1 - \frac{|i-j|}{k+1}\right)_+$$

- "Exponential" filter: positive definite but not banded

$$R_\sigma(i,j) = e^{-\frac{|i-j|}{\sigma}} = \rho^{|i-j|}$$

Thresholding

- $M \equiv ||m_{ij}||$
- $T_t(M) \equiv ||m_{ij}\mathbf{1}(|m_{ij}| \geq t)||$
  - Not positive definite in general
  - permutation invariant

The importance of l_2 operator norm analysis

- Matrix norms

$$
\begin{aligned}
M &\equiv \|m_{ij}\|_{p\times p} \\
|\mathbf{x}|_r^r &\equiv \sum_{j=1}^{p} |x_j|^r,\ \mathbf{x} = (x_1, \ldots, x_p)
\end{aligned}
$$

- Operator norms

$$
\begin{aligned}
\|M\|_{(r,s)} &\equiv \max\left\{ \frac{|M\mathbf{x}|_s}{|\mathbf{x}|_r} : \mathbf{x} \neq \mathbf{0} \right\} \\
\|M\|_{(2,2)} &= \lambda_{\max}^{1/2}(MM^T) \\
\|M\|_{(1,1)} &= \max_j \sum_{i=1}^{p} |m_{ij}| \\
\|M\|_{(\infty,\infty)} &= \max_i \sum_{j=1}^{p} |m_{ij}|
\end{aligned}
$$

The importance of l_2 operator norm analysis

- Other norms

$$\begin{aligned} \|M\|_\infty &\equiv \max_{i,j} |m_{ij}| \\ \|M\|_2^2 &\equiv \sum_{i,j} m_{ij}^2 \ : \ \text{Frobenius norm} \end{aligned}$$

- Which to use?
  - $\|M\|_\infty$: Easiest.

    But doesn't imply eigenstructures of inverses close.
  - $\|M\|_{(2,2)}$ Does but hard to analyze.
  - $\|M\|_2 \geq \|M\|_{(2,2)}$ But too big.

    $\|J\|_2 = p, \ \|J\|_{(2,2)} = 1.$

Properties

- For any operator norm,

$$||AB|| \leq ||A||||B||$$

  Henceforth, $||M||_{(2,2)} \equiv ||M||$.

- If $M_{p \times p}$ is symmetric,

$$||M|| = Max\left\{\left|\lambda_{Max}^{(M)}\right|, \left|\lambda_{Min}^{(M)}\right|\right\}.$$

- $||M|| \leq \left[||M||_{(1,1)}||M||_{(\infty,\infty)}\right]^{1/2}$.

  If M is symmetric, $||M|| \leq ||M||_{(1,1)}$.

Basic results I

Given A_n, B_n symmetric $||A_n - B_n|| \to 0$,

suppose $\lambda_1(B_n) > \lambda_2(B_n) > \cdots > \lambda_k(B_n) > \lambda_{k+1}(B_n)$

and define $\lambda_j(A_n)$ analogously.

Suppose $\lambda_{j+1}(B_n) < \lambda_j(B_n) - \triangle, \; 1 \leq j \leq k$

Dimension B_n arbitrary, $k, \triangle > 0$ fixed.

Then,

a) $|\lambda_j(A_n) - \lambda_j(B_n)| = O(\triangle^{-(j-1)}||A_n - B_n||)$

b) If E_{jA} respectively E_{jB} is projection operator onto eigenspace corresponding to λ_j, then

$$||E_{jA} - E_{jB}|| = O(\triangle^{-j}||A_n - B_n||)$$

Basic results II

NB:

$$\begin{aligned} \text{If } \lambda_j(A_n) &\leftrightarrow \mathbf{e}_{jnA} \ \text{ Eigenvector} \\ B_n &\leftrightarrow \mathbf{e}_{jnB} \end{aligned}$$

We have

$$|\mathbf{e}_{jnA} - \mathbf{e}_{jnB}| = O(\triangle^{-j}||A_n - B_n||)$$

B-L (2006) Main Result I

Banded estimator :

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

Let

$$\begin{aligned} \mathcal{U}(\epsilon_0, \alpha, C) &= \{\Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0, \\ & \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Ck^{-\alpha} \text{ for all } k \geq 0\}. \end{aligned}$$

Theorem 1

If $\mathbf{X}$ is Gaussian and $k_n \asymp \left(n^{-1}\log p\right)^{-\frac{1}{2(\alpha+1)}}$, then, uniformly on $\Sigma \in \mathcal{U}(\varepsilon_0, \alpha, C)$,

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

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

Remark

$\lambda_{min} \geq \varepsilon_0$ not needed if only convergence in $||\cdot||$ to Σ is needed.

Choosing the "banding" parameter

Ideally want to minimize risk

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

Estimate via **a resampling scheme**:

- Split the data into two samples of size n_1, n_2, N times at random
- Let $\hat{\Sigma}_1^{(\nu)}$, $\hat{\Sigma}_2^{(\nu)}$ be the two sample covariance matrices from the ν-th split. The risk can be estimated by

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

- We used $n_1 = n/3$, $N = 50$, and the L_1 matrix norm instead of L_2.

Simulation examples: banding $\hat{\Sigma}$

- Tridiagonal Σ (covariance of MA(1)): always pick $k = 1$.
- Covariance of AR(1): $\Sigma \in \mathcal{U}$

$$\sigma_{ij} = \rho^{|i-j|}$$

$n = 100,\ p = 10,\ 100,\ 200,\ \rho = 0.1,\ 0.5,\ 0.9.$

- Fractional Gaussian noise (FGN): long-range dependence, not in $\mathcal{U}$

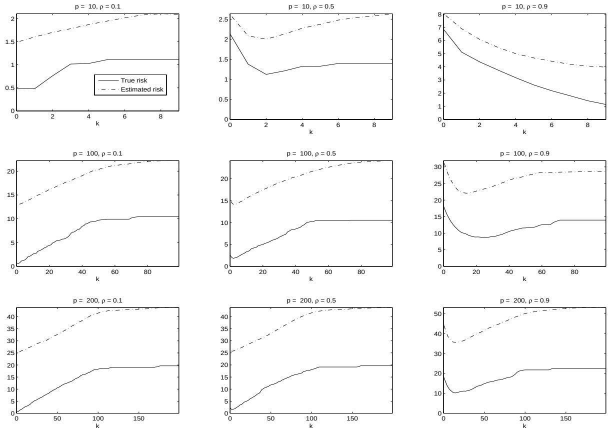
$$\sigma_{ij} = \frac{1}{2}\left[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right]$$

$H \in [0.5, 1]$ is the Hurst parameter

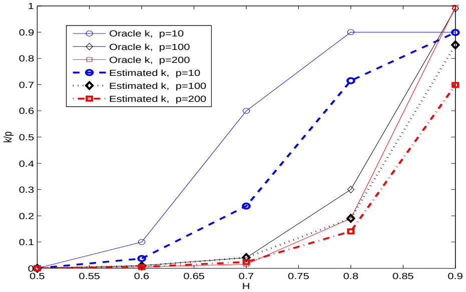
$H = 0.5$ is white noise; $H = 1$ is perfect dependence

$n = 100,\ p = 10,\ 100,\ 200,\ H = 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9.$

True and estimated risk for AR(1)



Ratio of optimal k to p for FGN

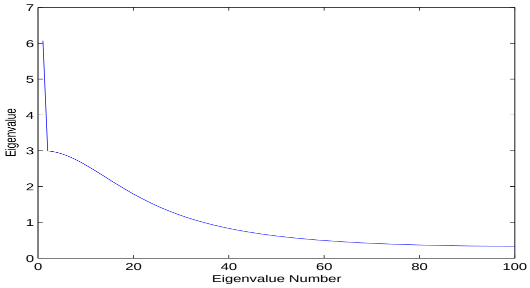


- The optimal amount of regularization is model dependent
- The same model requires more regularization in higher dimensions

Effect of banding on PCA

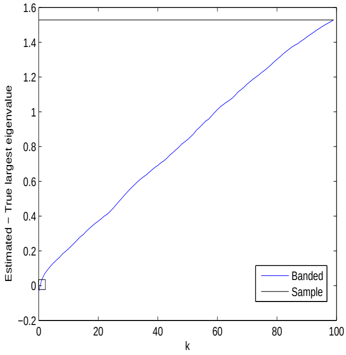
- Model: $X_i \sim \mathcal{N}_p(0, \Sigma)$, $n = 100$, $p = 100$.
- $\Sigma = \Sigma_0 + \mathrm{diag}(2\lambda_{\max}(\Sigma_0), 0, \ldots, 0)$, $[\Sigma_0]_{ij} = \rho^{|i-j|}$, $\rho = 0.5$

True eigenvalues

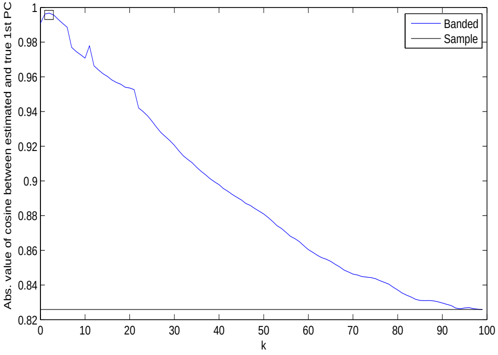


Estimation results for 1st principal component

$\hat{\lambda}_1 - \lambda_1$



$|\cos(\hat{e}_1, e_1)|$



- Resampling procedure picks $k = 2$.

El Karoui (2007) in progress

- Given Σ_p, $p \times p$ covariance matrix, compute adjacency matrix $A_p = \mathbf{1}_{\sigma(i,j)\neq 0}$
- Associate graph $\mathcal{G}_p$ to it
- Consider $\mathcal{C}_p(k) = \{$closed paths of length k on the graph with adjacency matrix $A_p\}$ and $\phi_p(k) = |\mathcal{C}_p(k)| = \mathrm{trace}\left(A_p^k\right)$.

Call sequence of Σ_p β-sparse if

$$\forall k \in 2\mathbb{N}\,, \phi_p(k) \leq f(k) p^{\beta(k-1)+1}$$

where $f(k)$ independent of p and $0 \leq \beta < 1$

Connection between closed paths and trace(Σ_p^k)

Examples : Computation of sparsity coefficients

- **Diagonal matrix** : $A_p = \mathrm{Id}_p$. $\phi(k) = p$, for all k. Sparsity coefficient: **0**.
- **Matrices with at most M non-zero elements on each line** $\phi(k) \leq pM^{k-1}$. Sparsity coefficient: **0**.
- **Matrices with at most Mp^{α} non-zero elements on each line** $\phi(k) \leq M^{(k-1)}pp^{\alpha(k-1)}$. Sparsity coefficient: $\boldsymbol{\alpha}$

Assumptions underlying results

In all that follows,

- $\Sigma_p(i,i)$ stay bounded
- $X_{i,j}$ have infinitely many moments
- Rows of $(n \times p)$ data matrix $\mathbf{X}$ i.i.d
- $p/n \to l \in (0, \infty)$

Simple case: gap in entries of covariance matrix

Gaussian MLE, centered case

- Suppose Σ_p β-sparse, $\beta = 1/2 - \eta$ and $\eta > 0$
- if $\sigma(i,j) \neq 0$, $|\sigma(i,j)| > Cn^{-\alpha_0}$, $0 < \alpha_0 = 1/2 - \delta_0 < 1/2$
- $X_{i,j}$ centered

Theorem

Let

$$S_p = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}'\mathbf{X}$$

$T_\alpha(S_p)$= thresholded version of S_p at level $Cn^{-\alpha}$ with $\alpha = 1/2 - \delta > \alpha_0$. Then,

$$|||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 \text{ a.s .}$$

Beyond truly sparse matrices

Approximation by sparse matrices

How does thresholding perform on matrices approximated by sparse matrices?

- Suppose $\exists T_{\alpha_1}(\Sigma_p) = \widetilde{\Sigma}_p$, β-sparse.
- Suppose $|||\widetilde{\Sigma}_p - \Sigma_p|||_2 \to 0$.
- Suppose $\exists \alpha_0 < \alpha_1 < 1/2 - \delta_0$ such that adjacency matrix of (i,j)'s such that $Cn^{-\alpha_1} < |\sigma(i,j)| < Cn^{-\alpha_0}$ is γ-sparse, $\gamma \leq \alpha_0 - \zeta_0$, $\zeta_0 > 0$.

proposition

Then conclusions of all the theorems above apply: for $\alpha \in (\alpha_0, \alpha_1)$,

$$|||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 \text{ a.s .}$$

A review of methods: some practiced I

1. $\widetilde{\Sigma} = \hat{\alpha}\widehat{\Sigma} + \hat{\beta}J$, (Ledoit, Wolf (2003))
   - Reasonable in some practice
   - No good theory
   - Useless for eigenstructure

2.
   (i) Sparse PCA, (Johnstone, Lu (2006))

   (ii) Supervised PCA, (Baird, Hastie, Paul, Tibshirani (2007)), (Paul (2007))

   To be discussed(?)

A review of methods: some practiced II

3. Regularizing the inverse

   (i) "Banding"

      a) Wu, Pourahmadi (2003)

      b) Bickel, Levina (2007)

   (ii) "Lasso"

      a) Huang, Liu,Pourahmadi, Liu (2006)

      b) M. Yuan (2007)

(i),(ii)

   a) Not permutation invariant

   b) $\hat{T} = \operatorname*{argmin}_T tr(T\widehat{\Sigma}) - \log(det(T)) + 2\lambda \sum_{i \neq j} |t_{ij}|$

A review of methods: some practiced III

4. Graphical models

- Meinshausen and Buhlmann (2006)
- Meinshausen and Yu (2006)
- Wainwright (2006)
- Kalisch and Buhlmann (2007)
- Zhao and Yu (2006)

Some important directions

- Choice of k or other regularization parameters.
- Canonical correlation regularized estimation.
- Independent component analysis regularization.
- Estimation of parameters governing independence and conditional independence in graphical models.

Some very in progress results I

Bickel and Levina (2007)

Theorem 1

Suppose $\Sigma \in \left\{ ||\sigma_{ij}|| : \max_j \sum_i |\sigma_{ij}|^q \leq C \right\}, 0 \leq q < 1.$

Then, in the Gaussian case, if $t_n \asymp M\sqrt{\frac{\log p}{n}}$,

$$\left\| T_{t_n}\left(\hat{\Sigma}\right) - \Sigma \right\| \asymp \left(\frac{\log p}{n}\right)^{(1-q)/2}$$

Some very in progress results II

theorem 2

Let $\Sigma \in \left\{ ||\sigma_{ij}|| : 0 < \varepsilon_0 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq \varepsilon_0^{-1},\ \sum_{i \neq j} \mathbf{1}(\sigma_{ij} \neq 0) \leq s \right\}$.

$$
\begin{aligned}
R &\equiv \left|\left|\sigma_{ij} \left[\sigma_{ii}\sigma_{jj}\right]^{-1/2}\right|\right| \\
S &\equiv Diag\left(\sigma_{ii}\right) \\
\hat{R} &\equiv \left|\left|\hat{\sigma}_{ij} \left[\hat{\sigma}_{ii}\hat{\sigma}_{jj}\right]^{-1/2}\right|\right| \\
\hat{S} &\equiv Diag\left(\hat{\sigma}_{ii}\right) \\
\tilde{\Sigma}^{-1} &= \hat{S}^{-1}\tilde{T}\hat{S}^{-1} \\
\tilde{T} &= \operatorname{argmin}\left\{ tr(\hat{R}T) - \log|T| + \lambda \sum_{i \neq j} |t_{ij}| \right\}
\end{aligned}
$$

Then, in the Gaussian or SubGaussian case,

$$
\left|\left|\tilde{T} - \Sigma^{-1}\right|\right| = O_p\left(\left(\frac{s \log p}{n}\right)^{1/2}\right)
$$

NB : s can be as large as $\binom{p}{2}$