Regularized Estimation of High Dimensional Covariance Matrices

Peter Bickel

Cambridge

January, 2008

With Thanks to

- E. Levina (Joint collaboration, slides)
- I. M. Johnstone (Slides)
- Choongsoon Bae (Slides)

Outline

1. Why estimate covariance matrices?
2. Some examples
3. Pathologies of empirical covariance matrix
4. Sparsity in various guises
5. A brief review of methods
6. Asymptotic theory for banding, thresholding, SPICE (LASSO penalized Gaussian log likelihood)
7. Simulations and data
8. Some future directions

Model

$\mathbf{X}_1, \ldots, \mathbf{X}_n$: p vectors, I.I.D. F.

- Base model: $F = N_p(\boldsymbol{\mu}, \Sigma)$
- More generally: $\mathbb{E}_F|\mathbf{X}|^2 < \infty$

  $\boldsymbol{\mu}(F) = \mathbb{E}_F\mathbf{X}$

  $\Sigma(F) = \mathbb{E}_F(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T$
- Asymptotic theory: $n, p \to \infty$, $\Sigma \in \mathcal{T}_{p,n}$
- Extensions to Stationary $\mathbf{X}_1, \ldots, \mathbf{X}_n$, Wei Biao Wu (2007).

Why Estimate Covariance Matrix?

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

Covariance itself is usually not the end goal:

PCA: Estimation of the eigenstructure

LDA/QDA, Graphical models: Estimation of inverses

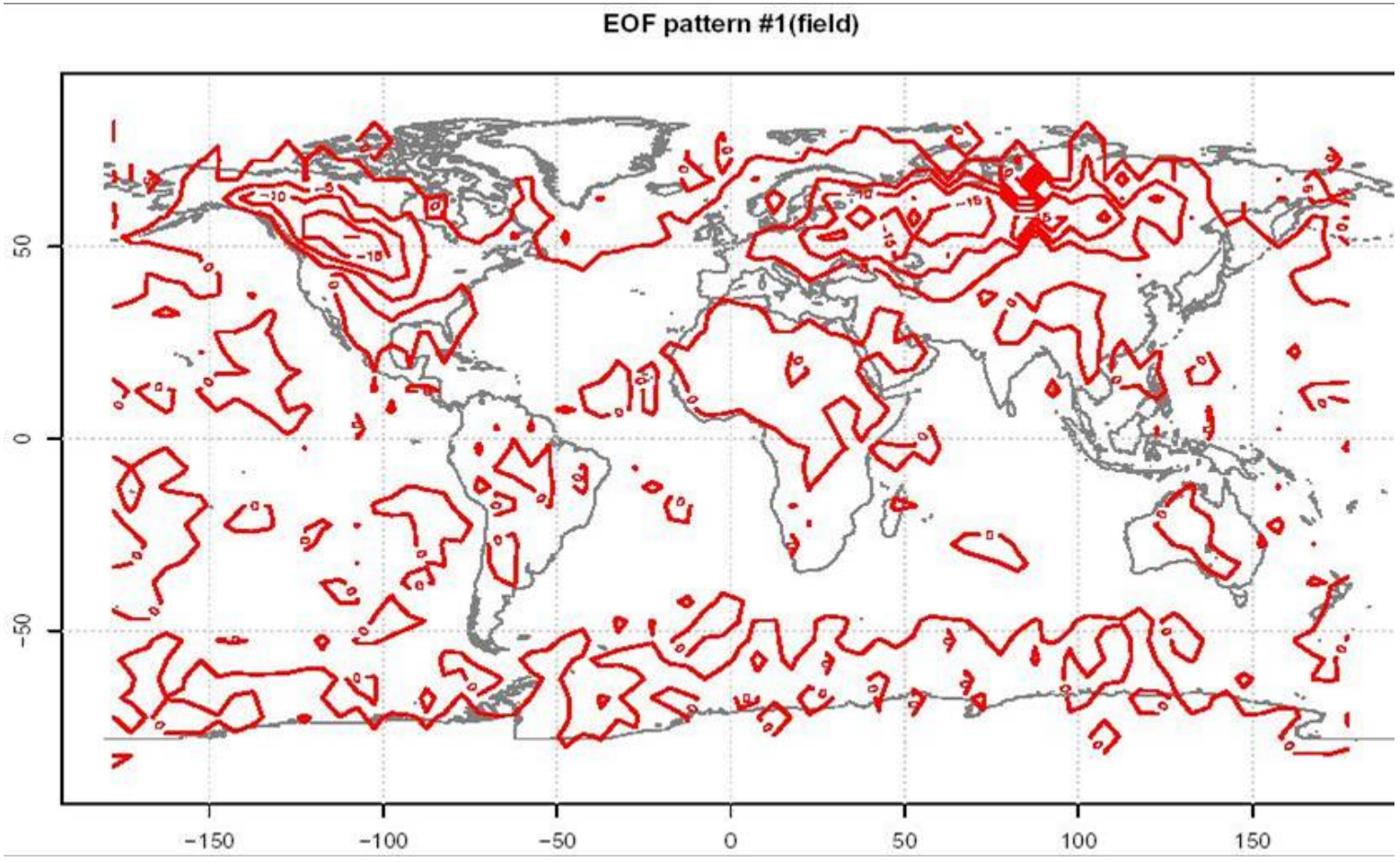
Covariance Matrices in Climate Research

- Example 1

  Data courtesy of Serge Guillas

  - Empirical Orthogonal Functions(EOF) aka principal components.
  - $\mathbf{X}_1, \ldots, \mathbf{X}_n$: p dimensional stationary vector time series.
  - Monthly Jan Temp $1850 - 2006$
  - Latitude longitude calls $5^o \times 5^o$ grid
  - p=2592
  - n=157

Example 1



EOF pattern #1 (clim.pact)

Example 1

- $\mathbf{X}_k = \{X(i,j) \ : \ i = \text{latitude}, \ j = \text{longitude}\}$
- $\mathbb{X}_{n\times p} = \left(\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T\right)^T$
- $\mathbb{E}(\mathbf{X}) = \mathbf{0}$
- $\Sigma = Var(\mathbf{X}) = \mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right) = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j^T$
- $\mathbf{e}_1, \ldots, \mathbf{e}_p$: 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

- Example 2

Data assimilation

  - $X_j =$ ave. "pressure", "temperature",... in $50km \times 50km \times$ *variable* block of atmosphere, $|J| \asymp 10^7$, computer model.
  - $\mathbf{X}_i = \mathbf{X}(t_i)\ i = 1, \ldots, T.$
  - $\mathbf{Y}(t_i)$: 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

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

- 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 can be inconsistent if $p/n \to \gamma > 0$.

Eigenvalues

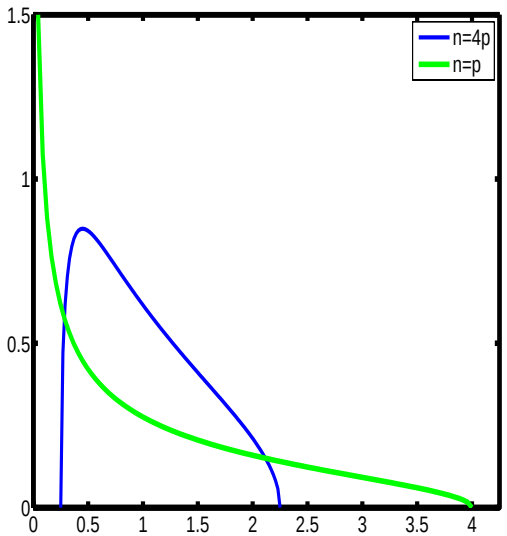
Description of spreading phenomenon:

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

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

Marčenko-Pastur, (67), For $F = N_p(\mathbf{0}, I)$, $\quad p/n \to \gamma$

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



Eigenvectors

D.Paul (2006) For $N_p(\mathbf{0}, \Sigma)$

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

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

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

Fisher's LDA

- For classification into two populations using:

  $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N_p(-\triangle, \Sigma)$, $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{2n} \sim N_p(\triangle, \Sigma)$

- Minimax behaviour over $\mathcal{T}_{p,n} = \left\{\Sigma : \triangle^T \Sigma^{-1} \triangle \geq c > 0\right\}$ of LDA based on MLE $\widehat{\Sigma}^{-1}$ (Moore-Penrose for $p \geq 2^{n-2}$) is equivalent to coin tossing if $\frac{p}{n} \to \infty$ (Bickel - Levina (2004))

- True for any rule!. What if $\mathcal{T}$ consists of "sparse" matrices?

Eigenstructure Sparsity

- Write,

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

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

- Equivalent :

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

Eigenstructure Sparsity defined

- 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 $\mathbf{v}$.

Label Dependent Sparsity of Σ

- Write $\Sigma = ||\sigma_{ij}||$,
  - $|\sigma_{ij}|$ small (effectively 0) for $|i - j|$ large
  - More generally, given metric m on J, $|\sigma_{ij}|$ small if $m(i, j)$ large
- Note: $\sigma_{ij} = 0 \Leftrightarrow X_i \perp X_j$ under Gaussianity.

Label Dependent Sparsity of Σ^{-1}

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

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

Note:

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

Examples of Label Dependent Sparsity

(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:

- $\sum_{i,j} K^2(i,j) < \infty$ ($Z_m \xrightarrow{p} 0$, non stationary)
- $(i) \Rightarrow \sum_{i} s^2(i) < \infty$
- $S = 0 \Rightarrow \Sigma$ singular but correspond to functional data analysis

A typical goal: Estimate Σ^{-1} for prediction.

Permutation Invariant Sparsity of Σ *or* Σ^{-1}

a) Each row of Σ sparse or sparsely approximable

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

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

- *a*) roughly implies *b*) if $\lambda_{min}(\Sigma) \geq \delta > 0$.
- A more general notion (El Karoui (2007) AS to appear).

(Roughly) The number of loops of length k in the undirected graph:

$i \sim j$ iff $\sigma_{ij} \neq 0$ grows more slowly than $p^{k/2}$ as $k \to \infty$. The corresponding definition for Σ^{-1} is not equivalent in general.

Permutation Invariant Sparsity of Σ^{-1}

- Σ^{-1} corresponds to a graphical model.
- $\mathcal{N}(i) = \{j : \sigma^{ij} \neq 0, j \neq i\}$
- Sparsity means neighbourhood size $\ll p$.

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

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Eigenstructure sparsity**
  - *a)* Sparse PCA d'Aspremont *et al.*, SIAM Review (2007), Hastie, Tibshirani, Zhou (2007).
    - PCA with Lasso penalty on coordinates of eigenvectors
    - Focus on computation, no asymptotics

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Eigenstructure sparsity continued**

  b) Johnstone and Lu, JASA (2006).

  - Sparse PCA for Σ corresponding to signals $\mathbf{X} = (X(t_1), \ldots, X(t_p))^T$
  - Represent in wavelet basis
  - Use only $k << p$ coefficients to represent data to whose covariance matrix PCA is applied
  - Asymptotics for large p and examples

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Label Dependent Sparsity**
  - Banding

Bickel and Levina (2004), Bernoulli

Bickel and Levina (2007), Annals of Statistics (To appear)

a) "Banding" $\widehat{\Sigma}$

For any $M \equiv ||m_{ij}||$,

$$\begin{aligned} B_k(M) &\equiv ||m_{ij}\mathbf{1}(|i-j| \leq k)|| \\ \widehat{\Sigma} &\equiv \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \end{aligned}$$

Estimate: $B_k(\widehat{\Sigma})$, $k = o(p)$.

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Label Dependent Sparsity continued**

  b) "Banding" Σ^{-1}

$$\Sigma^{-1} = AA^T, \ A \text{ Lower triangular}$$

Estimate: $\widehat{\Sigma}^{-1} \equiv \hat{A}_k \hat{A}_k^T$, where $\hat{A}_k$ is obtained by regression.

Pourahmadi and Wu (2003), Biometrika.

Huang,Liu,Pourahmadi (2007)

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Permutation Invariant Sparsity**

  (a) Thresholding

  For any $M = ||m_{ij}||$, $T_t(M) \equiv ||m_{ij}\mathbf{1}(|m_{ij}| \geq t)||$

  Estimate: $T_t(\widehat{\Sigma})$, $|t| = o(1)$

  El Karoui (2007)

  Bickel and Levina (2007).

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Permutation Invariant Sparsity continued**

  b) SPICE

  Estimate of Σ^{-1}: Σ^{-1} sparse

$$F(P, M) \equiv Trace(MP) - \log det(P) + \lambda |P|_1,$$

  where $|M|_1 = \sum_{i,j} |m_{ij}|$

  a) $\hat{P}^{(1)} \equiv \arg\min \left\{ F(P, \hat{\Sigma}) : P \succeq 0 \right\}$

  Yuan, Li (2007), Biometrika. Fixed p asymptotics

  Banerjee *et al.* (2006), ICML computation

  b) $\hat{R}^{(2)} \equiv \arg\min \left\{ F(P, \hat{R}) : R \succeq 0, r_{ii} = 1, \ \ all\ i \right\}$

  $\hat{P}^{(2)} \equiv \hat{D}\hat{R}\hat{D}$, $\hat{D} \equiv diag(\hat{\sigma}_{ii})$.

  $\hat{R} \equiv \hat{D}^{-1}\hat{\Sigma}\hat{D}^{-1}$

  a),b): Rothman *et al.* (2007), Submitted to JRSS(B) Asymptotics

Construction of Estimates of Σ or Parameters of Σ which work well for $\mathcal{T}$ having sparse members

- **Permutation Invariant Sparsity continued**
  - *c)* Methods for neighbourhood estimation
    - Meinshausen and Buhlmann (2006), Zhao and Yu (2007): Regression methods + Lasso
    - Wainwright (2006): Bounds
    - Kalisch and Buhlmann (2007): Using the PC algorithm of Spertes *et al.* (2000) to roughly hierarchically test whether partial correlation coefficients are 0

Some Asymptotic Theory

- 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

$$
\begin{aligned}
\|M\|_{(2,2)} &= \lambda_{\max}^{1/2}(MM^T) \\
\|M\|_{(1,1)} &= \max_j \sum_{i=1}^{p} |m_{ij}| \\
\|M\|_{(\infty,\infty)} &= \max_i \sum_{j=1}^{p} |m_{ij}| \\
|M|_\infty &\equiv \max_{i,j} |m_{ij}| \\
\|M\|_F^2 &\equiv \sum_{i,j} m_{ij}^2 \ : \ \text{Frobenius norm}
\end{aligned}
$$

Properties

- For any operator norm,

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

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

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

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

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

Basic Property of $||\cdot||$

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

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

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

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

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 $\Sigma \in \mathcal{U}(\varepsilon_0, \alpha, C)$,

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

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

Remark

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

An Approximation Theorem (Bickel and Lindner (2007))

If $m||B_k - A|| \leq \delta$, (B_k banded of width $2k$), $||A|| = M < \infty$, $||A^{-1}|| = m^{-1} < \infty$, $K \equiv \frac{M}{m}$,

There exists $N(m, M, \delta)$, B_{kN} such that

$$\left|\left|B_{kN} - A^{-1}\right|\right| \leq \frac{1}{M} c(m, M, \delta)\delta,$$

where $c(m, M, \delta)$ tends to $c(K)$ as $\delta \to 0$.

B_{kN} has a simply computable form.

Thresholding

Theorem *(Thresholding) (B-L Submitted to AS)*

If $\mathcal{U}_t(q, c_0(p), M) = \left\{ \Sigma : \sigma_{ii} \leq M_0, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_0(p), \ \textit{all } i \right\}$,

$0 \leq q < 1$, $t_n = M\sqrt{\dfrac{\log p}{n}}$,

$$\left|\left| T_{t_n}\left(\hat{\Sigma}\right) - \Sigma \right|\right| = O_p\left(c_0(p) \left(\frac{\log p}{n} \right)^{(1-q)/2} \right)$$

Note

$q = 0 \ \leftrightarrow \ c_0(p)$: non zero entries per row

More subtle consistency results : N. El Karoui (2007).

SPICE (Rothman et al.*)*

If $S \equiv \left\{(i,j) : \sigma^{ij} \neq 0\right\}$, $|S| \leq s$, $||\Sigma^{-1}|| \leq m^{-1}$, $||\Sigma|| \leq M$ and $\mathbf{X}$ is Gaussian,

$$\begin{aligned} \left|\left|\hat{P}^{(1)} - \Sigma\right|\right|_F^2 &= O_p\left((p+s)\frac{\log p}{n}\right) \\ \left|\left|\hat{P}^{(2)} - \Sigma\right|\right|^2 &= O_p\left(\frac{s \log p}{n}\right) \end{aligned}$$

Choosing the "Banding" (or Thresholding or SPICE) 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.
- Oracle properties for Frobenius norm (Bickel and Levina (2007), Annals of Statistics submitted).

Simulation Examples

- 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 $\mathcal{U}$

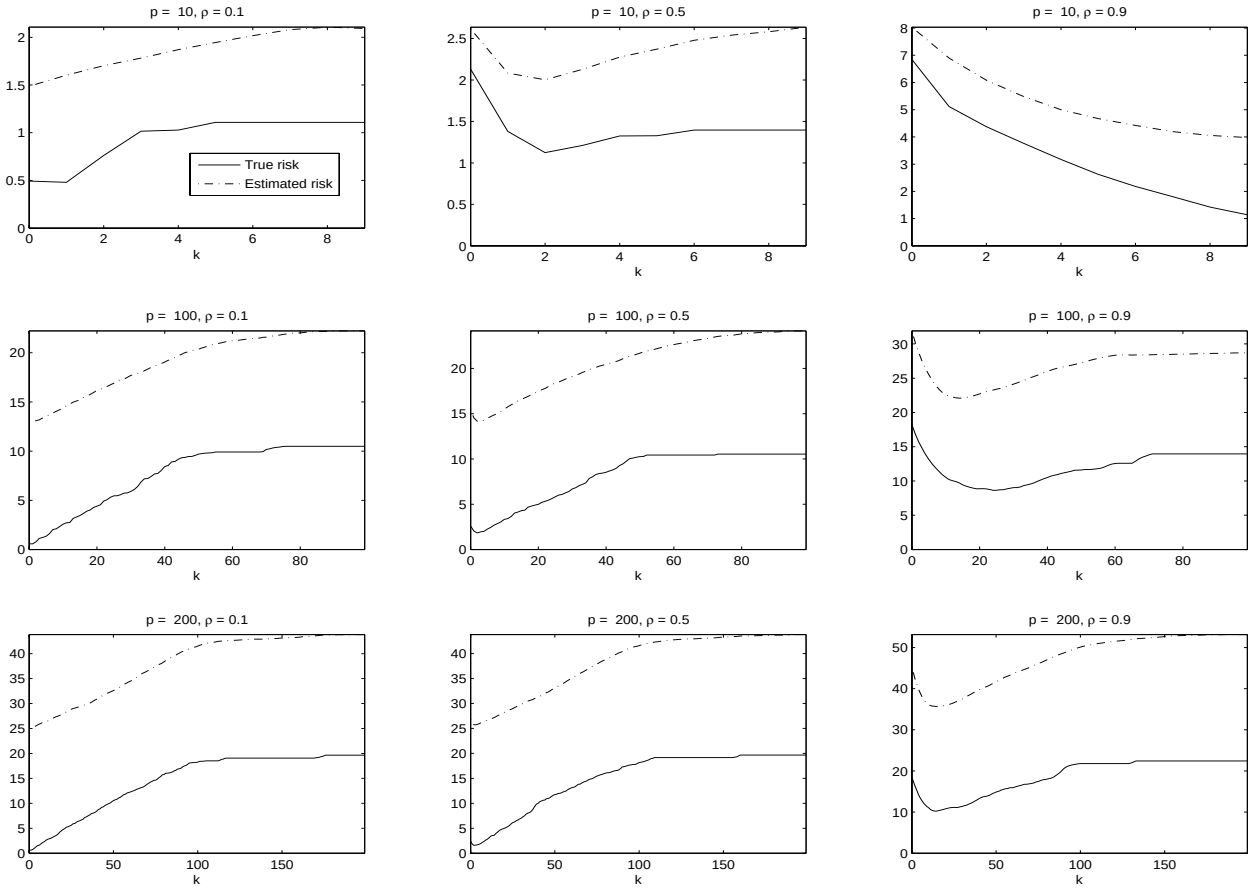
$$\sigma_{ij} = \frac{1}{2}\left[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right]$$

$H \in [0.5, 1]$ 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)



$(1,1)$ *Loss and Choice of k for AR(1)*

		Mean(SD)		Loss		
p	ρ	k_1	$\hat{k}$	$\hat{\Sigma}_{\hat{k}}$	$\hat{\Sigma}_{k_1}$	$\hat{\Sigma}$
10	0.1	0.5(0.5)	0.0(0.2)	0.5	0.4	1.1
10	0.5	3.3(0.8)	2.0(0.6)	1.1	1.0	1.3
10	0.9	8.6(0.7)	8.9(0.3)	1.5	1.5	1.5
100	0.1	0.2(0.4)	0.1(0.3)	0.6	0.6	10.2
100	0.5	2.7(0.7)	2.3(0.5)	1.6	1.5	10.6
100	0.9	21.3(4.5)	15.9(2.6)	9.2	8.5	13.5
200	0.1	0.2(0.4)	0.2(0.4)	0.7	0.6	20.4
200	0.5	2.4(0.7)	2.7(0.5)	1.8	1.7	20.8
200	0.9	20.2(4.5)	16.6(2.4)	9.9	9.5	24.5

Table: AR(1): Oracle and estimated k and the corresponding loss values.

$$k_1 = \underset{k}{\operatorname{argmin}} \|\hat{\Sigma}_k - \Sigma\|_{(1,1)}.$$

Operator Norm Results for Various Estimates and Situations

p	ρ	Sample	Band	Band-P	Thresh	SPICE
10	0.1	0.60(0.01)	0.35(0.01)	0.35(0.01)	0.40(0.01)	0.35(0.01)
10	0.5	0.73(0.02)	0.64(0.02)	0.74(0.02)	0.80(0.02)	0.72(0.02)
10	0.9	1.02(0.05)	1.02(0.05)	1.02(0.05)	1.02(0.05)	1.01(0.05)
100	0.1	2.86(0.02)	0.45(0.01)	0.45(0.01)	0.50(0.01)	0.45(0.01)
100	0.5	3.39(0.03)	0.95(0.01)	2.04(0.00)	1.96(0.01)	1.52(0.01)
100	0.9	6.03(0.12)	4.45(0.09)	6.03(0.12)	6.18(0.12)	5.57(0.12)
200	0.1	4.63(0.02)	0.49(0.01)	0.49(0.01)	0.51(0.01)	0.50(0.01)
200	0.5	5.47(0.03)	1.00(0.01)	2.06(0.00)	2.04(0.00)	1.60(0.01)
200	0.9	9.77(0.16)	5.24(0.09)	9.76(0.16)	8.80(0.10)	7.50(0.10)

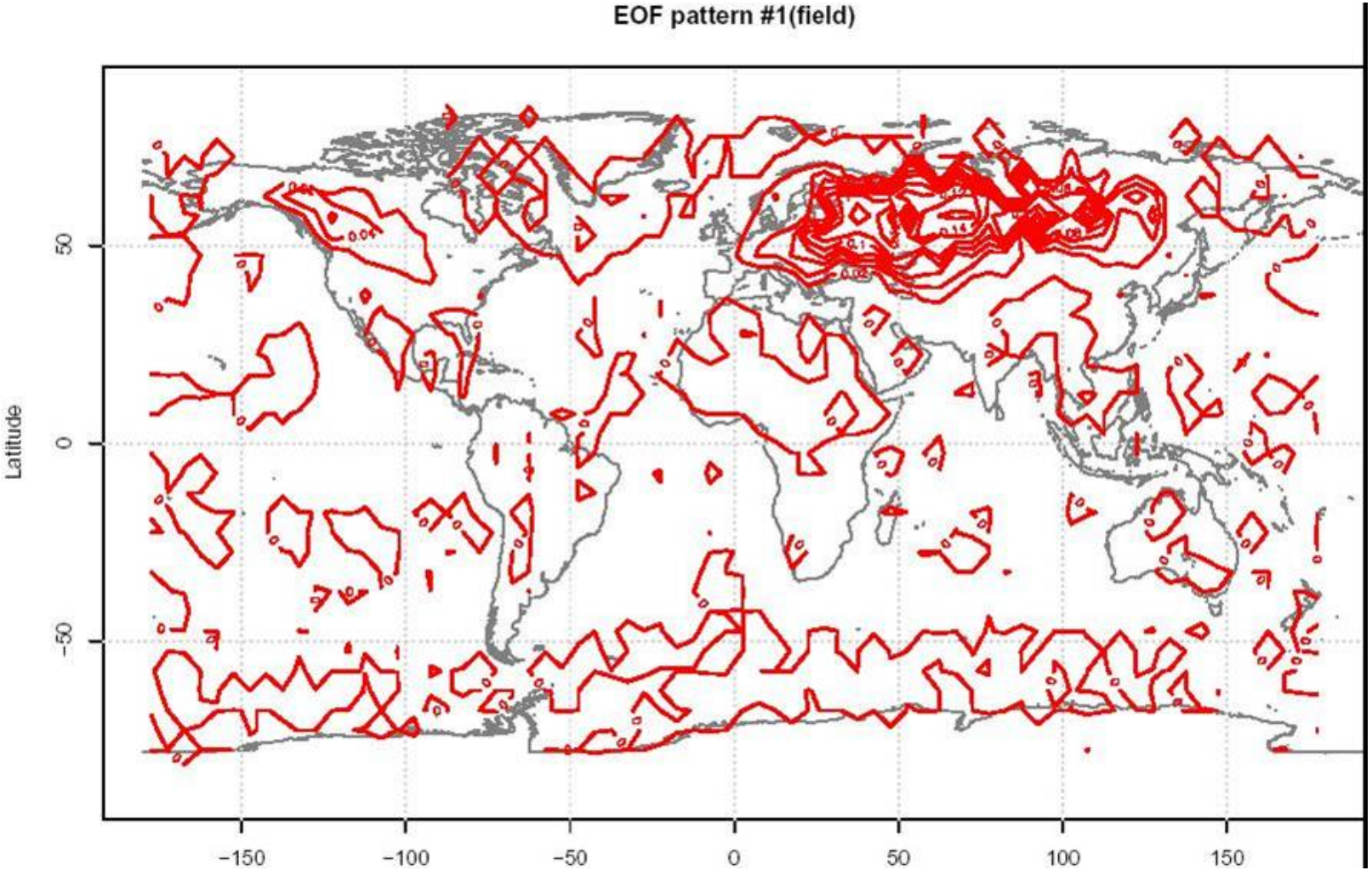
Table: Operator Norm Loss, Averages and SEs over 100 replications. AR(1)

Operator Norm Results for Various Estimates and Situations

p	H	Sample	Band	Band-P	Thresh	SPICE
10	0.5	0.60(0.01)	0.27(0.01)	0.27(0.01)	0.35(0.01)	0.28(0.01)
10	0.7	0.67(0.02)	0.88(0.02)	1.03(0.03)	0.88(0.04)	1.05(0.05)
10	0.9	0.96(0.04)	0.96(0.04)	0.96(0.04)	0.96(0.04)	1.13(0.05)
100	0.5	2.83(0.01)	0.38(0.01)	0.38(0.01)	0.46(0.01)	0.39(0.01)
100	0.7	3.43(0.05)	4.46(0.02)	5.37(0.00)	5.36(0.00)	5.30(0.02)
100	0.9	8.02(0.28)	8.03(0.28)	8.02(0.28)	8.02(0.28)	14.18(0.70)
200	0.5	4.60(0.02)	0.44(0.01)	0.44(0.01)	0.46(0.01)	0.45(0.01)
200	0.7	5.81(0.06)	6.46(0.01)	7.40(0.00)	7.40(0.00)	7.39(0.00)
200	0.9	14.34(0.43)	14.65(0.46)	14.34(0.43)	14.34(0.43)	30.16(0.60)

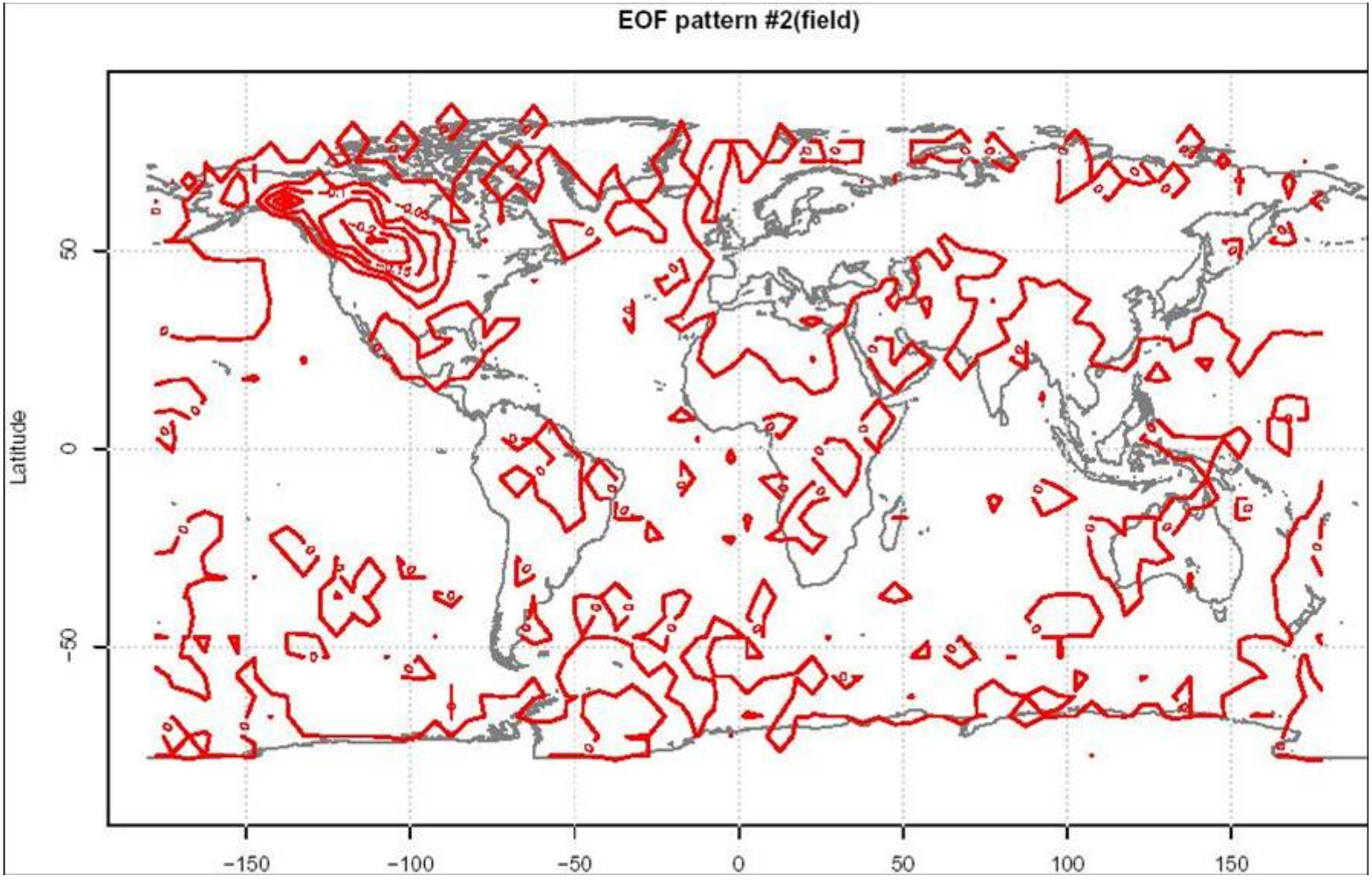
Table: Operator Norm Loss, Averages and SEs over 100 replications. FGN

Example 1 continued



EOF pattern #1 (Thresholding)

Example 1 continued



EOF pattern #2 (Thresholding)

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.