

# Estimation of large covariance matrices

**Liza Levina**

Department of Statistics  
University of Michigan

Part joint work with **Peter Bickel** (UC Berkeley)

Part joint work with **Ji Zhu** (University of Michigan)

## Why estimate covariance?

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

Covariance itself is usually not the end goal:

- PCA requires estimation of the **eigenstructure**
- LDA/QDA and conditional independence require the **inverse**

## What's wrong with the sample covariance matrix?

Observe  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , 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, unbiased (almost), well-behaved (and well studied) for fixed  $p$ ,  $n \rightarrow \infty$ . But very **noisy** if  $p$  is large.
- Eigenvalues overdispersed [[Marcenko-Pastur \(1967\)](#), [Wachter \(1978\)](#), [Geman \(1980\)](#), [Bai and Yin \(1993\)](#), [Johnstone \(2001\)](#), [Paul \(2004\)](#)]
- Eigenvectors are not consistent ([Johnstone and Lu, 2004](#))
- LDA breaks down if  $p/n \rightarrow \infty$  ([Bickel and Levina, 2004](#))
- **Singular** if  $p > n$

# Alternatives to the sample covariance matrix

## I. Steinian shrinking of sample eigenvalues

- First proposed by Stein (Rietz lecture, 1975)
- Empirical Bayes (Haff, 1980):  $\rho_1 \hat{\Sigma} + \rho_2 I$ ,  $\rho_1, \rho_2$  depend on  $p, n$  only
- Minimax shrinkage (Stein, Dey and Srinivasan (1985)): adjusted eigenvalues are neither positive nor ordered
- Ledoit and Wolf (2003):  $\rho_1 \hat{\Sigma} + \rho_2 I$ , optimal  $\rho_1, \rho_2$  estimated from data
- The form  $\rho_1 \hat{\Sigma} + \rho_2 I$  also used in other contexts:
  - original formulation of ridge regression (Hoerl and Kennard, 1970),
  - regularized discriminant analysis (Friedman, 1989)

All these do not change eigenvectors.

## II. Cholesky decomposition

Pourahmadi (1999): longitudinal data

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

$$\hat{X}_j = \sum_{l=1}^{j-1} \phi_{jlt} X_t, \quad \varepsilon_j = X_j - \hat{X}_j, \quad d_j^2 = \text{Var}(\varepsilon_j)$$

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

$$\varepsilon = \mathbf{X} - \hat{\mathbf{X}} = T\mathbf{X}$$

Independence of residuals  $\Rightarrow$  modified Cholesky decomposition:

$$D = T\Sigma T^T, \quad \Sigma^{-1} = T^T D^{-1} T$$

- Transforms covariance estimation into a regression problem

## Estimators based on Cholesky decomposition

- Are always positive definite ( $\Sigma^{-1} = T^T D^{-1} T$ )
- Shrink elements of  $T$
- **Not invariant under variable permutations**  $\Rightarrow$  most appropriate when there is a **natural ordering in the data** (e.g., time series, longitudinal data, spectroscopy, etc)
- Implicitly assume  $|i - j|$  large implies  $X_i$  and  $X_j$  nearly independent given the intervening variables
- Become singular if  $p > n$  unless regularized
- Give a natural estimate of  $\Sigma^{-1}$  rather than  $\Sigma$

- Wu and Pourahmadi (2003)
  - **Banding**: smooth the first  $k$  sub-diagonals with a spline, set the rest to 0. Choose  $k$  by AIC/BIC.
  - Element-wise convergence with rates determined by splines
- Huang, Liu, Pourahmadi, and Liu (2006)
  - Fit  $T$  and  $D$  by **maximum likelihood** with lasso or ridge **penalty**
- Levina and Zhu (2006...in progress)
  - **Adaptive banding** with a **hierarchical lasso** penalty

### III. Regularization by banding or tapering

- 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

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

- Banding Toeplitz matrices (Bickel and Levina, 2004):
  - leads to convergence to Bayes risk for LDA
  - not evaluated in the context of general estimation
- Tapering (Furrer and Bengtsson, 2006):
  - Tapering covariance in the context of Kalman filtering
  - $R$  is a function of  $\hat{\Sigma}$
  - Some convergence results in Frobenius norm
- Also not invariant under permutations (need a natural ordering)
- Implicitly assume  $|i - j|$  large implies  $X_i$  and  $X_j$  nearly uncorrelated

**There are other estimators...**

## Convergence of regularized estimators

Bickel and Levina (2006)

- All results in **operator norm**, a.k.a. **the matrix  $L_2$  norm**: for symmetric positive definite  $M$ ,

$$\|M\| = \lambda_{\max}(M)$$

- Results uniform over classes of covariance matrices as  $p, n \rightarrow \infty$

**Banding the covariance matrix:** define the class

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

Includes stationary processes with bounded smooth spectral density + well-behaved non-stationary noise

## Main Result I

Banded estimator:

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

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

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

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

## Extensions

- Gaussianity may be replaced by

$$P(X_j^2 > t) \leq C e^{-\gamma t} \quad \text{for all } j$$

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

$$R_\sigma(a, b) = g\left(\frac{\rho(a, b)}{\sigma}\right)$$

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

– includes triangular and exponential filters

- Also show that under the “spike” model, the estimated top eigenvalues and the corresponding eigenvectors are close to the truth

## Banding the Cholesky factor

- Center variables
- Regress  $X_j$  on  $X_{j-1}, \dots, X_{j-k}$ ; get new matrices of coefficients  $\tilde{A}_k$ , and residual variances  $\tilde{D}_k$ ; define  $\tilde{T}_k = I - \tilde{A}_k$ , and let

$$\begin{aligned}\tilde{\Sigma}_{k,p}^{-1} &= \tilde{T}_k^T \tilde{D}_k^{-1} \tilde{T}_k, \\ \tilde{\Sigma}_{k,p} &= \tilde{T}_k^{-1} \tilde{D}_k [\tilde{T}_k^{-1}]^T.\end{aligned}$$

- $\tilde{\Sigma}_k^{-1}$  is  $k$ -banded nonnegative definite;  $\tilde{\Sigma}_k$  is in general not banded, and different from  $\hat{\Sigma}_k$
- Similarly,  $\tilde{\Sigma}_k^{-1}$  is not the same as banded  $\hat{\Sigma}^{-1}$ , which is ill-defined when  $p > n$ .

## Main result II

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

$$\mathcal{U}^{-1}(\varepsilon_0, C, \alpha) = \left\{ \Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \varepsilon_0^{-1}, \right.$$

$$\left. \max_i \sum_{j < i-k} |t_{ij}(\Sigma)| \leq Ck^{-\alpha} \text{ for all } k \leq p-1 \right\}$$

**Theorem 2:** Uniformly for  $\Sigma \in \mathcal{U}^{-1}(\varepsilon_0, C, \alpha)$ , if  $\mathbf{X}$  is Gaussian,

$$k_n \asymp (n^{-1/2} \log p)^{-\frac{1}{\alpha+1}}, \text{ and } n^{-1/2} \log p = o_P(1),$$

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

## 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  $\nu$ -th split. The risk can be estimated by

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

- We used  $n_1 = n/3$ ,  $N = 50$ , and the  $L_1$  matrix norm instead of  $L_2$ .
- Same technique can be used for the Cholesky-based  $\tilde{\Sigma}_k$

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

- Tridiagonal  $\Sigma$  (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  $\mathcal{U}$

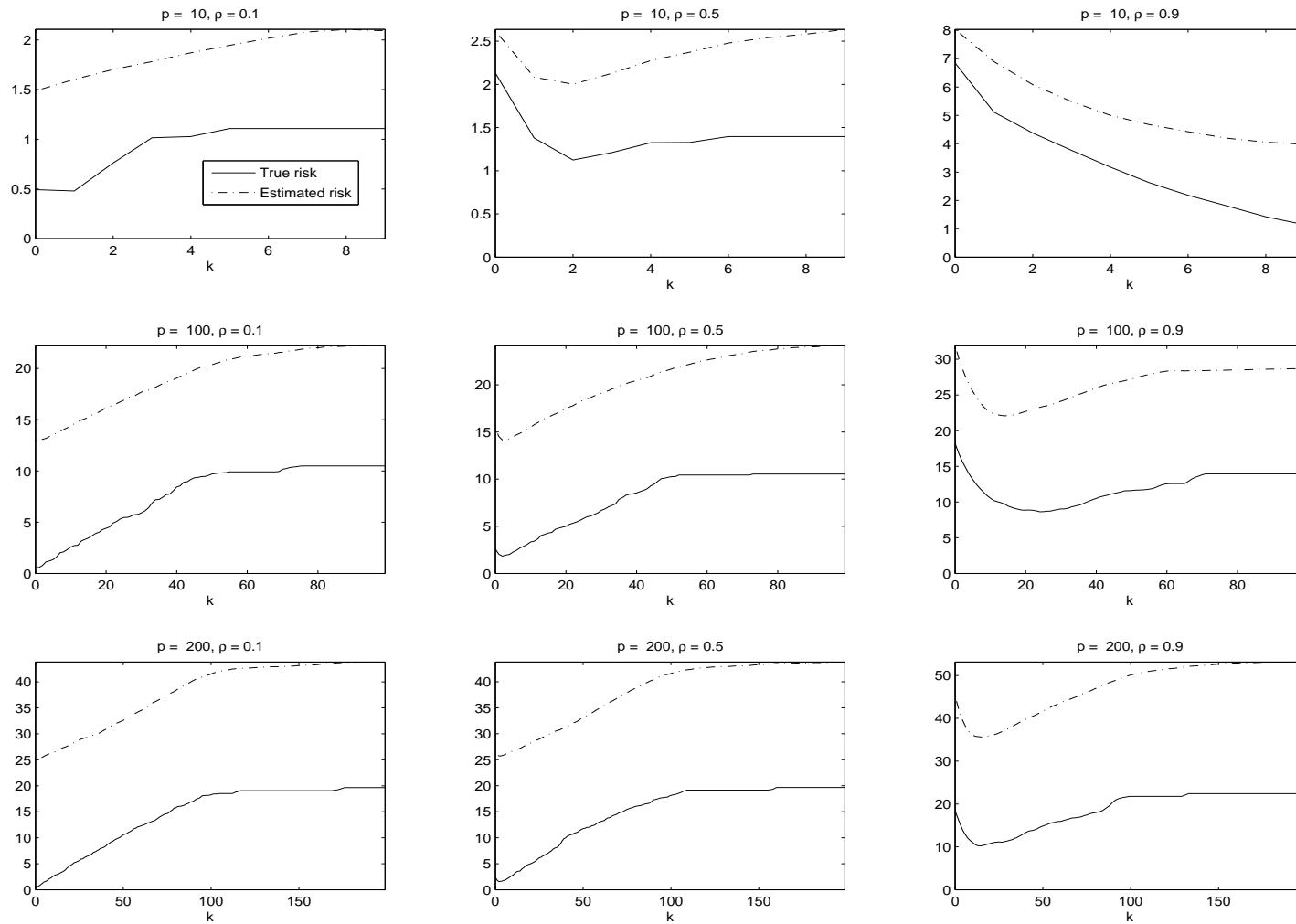
$$\sigma_{ij} = \frac{1}{2} [ (|i-j| + 1)^{2H} - 2|i-j|^{2H} + (|i-j| - 1)^{2H} ]$$

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

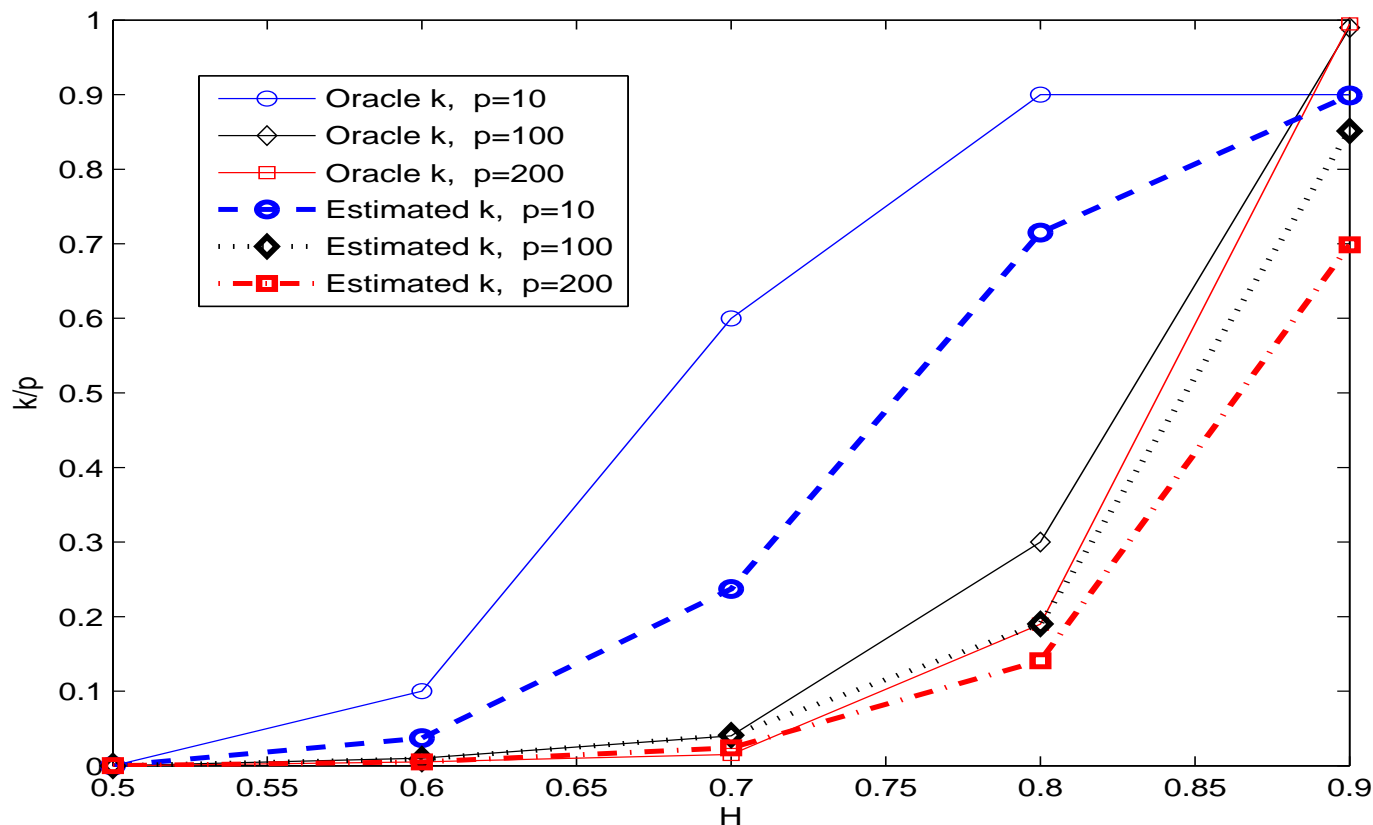
$H = 0.5$  is white noise;  $H = 1$  is perfect dependence

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

# 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**

## Adaptive banding of the Cholesky factor

(Levina and Zhu, 2006)

**Penalized likelihood** (Huang et al., 2006):

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

$$\begin{aligned}
 \ell(\mathbf{x}_1, \dots, \mathbf{x}_n, \Sigma) &= n \ln |\Sigma| + \sum_{i=1}^n \mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i \\
 &= \sum_{j=1}^p \ell_j(\mathbf{x}_1, \dots, \mathbf{x}_n, \Sigma) \\
 &= \ln d_1^2 + \frac{1}{d_1^2} \sum_{i=1}^n x_{i1}^2 + \\
 &+ \sum_{j=2}^p \left( \ln d_j^2 + \frac{1}{d_j^2} \sum_{i=1}^n \left( x_{ij} - \sum_{t=1}^{j-1} \phi_{jt} x_{it} \right)^2 \right)
 \end{aligned}$$

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):

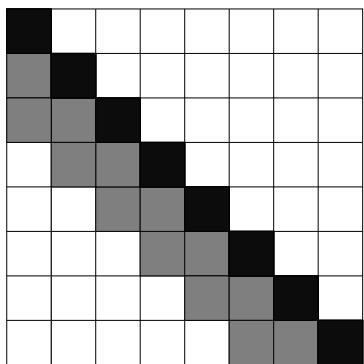
$$J(\phi_j) = \lambda \sum_{t=1}^{j-1} |\phi_{jt}|$$

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

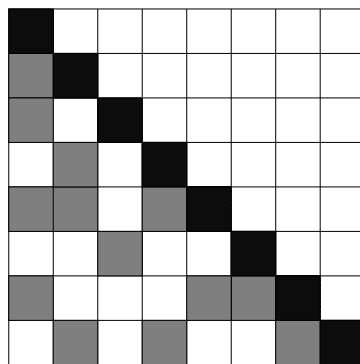
## Hierarchical LASSO penalty

$$J(\phi_j) = \lambda \left( |\phi_{j,j-1}| + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + \dots + \frac{|\phi_{j,1}|}{|\phi_{j,2}|} \right)$$

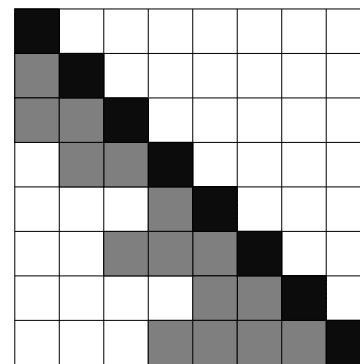
- Shrinkage + sparsity: If  $\hat{\phi}_{jt} = 0 \implies \hat{\phi}_{jj'} = 0$  for all  $j' < t$ .
- Sparse in  $T$  and  $\Sigma^{-1}$
- Hierarchical LASSO  $\implies$  Adaptive Banding of  $\Sigma^{-1}$



Banding



Lasso



Adaptive banding

- Scale could be an issue. Can replace  $J(\phi_j)$  by

$$J_1(\phi_j) = \lambda \left( \frac{|\phi_{j,j-1}|}{|\hat{\phi}_{j,j-1}^0|} + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + \dots + \frac{|\phi_{j,1}|}{|\phi_{j,2}|} \right)$$

$$J_2(\phi_j) = \lambda_1 \sum_{t=1}^{j-1} |\phi_{j,t}| + \lambda_2 \sum_{t=1}^{j-2} \frac{|\phi_{j,t}|}{|\phi_{j,t+1}|}$$

where  $\hat{\phi}_{j,j-1}^0$  is the coefficient from regressing  $X_j$  on  $X_{j-1}$  alone.

- When data is **not normal**, using normal likelihood may be misleading.  
Can instead fit each regression by **penalized least squares**.

## The algorithm

Iterate between 1 and 2 until convergence

1. Fix  $\phi_j$ , solve for  $d_j$  (easy)
2. Fix  $d_j$ , solve for  $\phi_j$ : iterative procedure
  - Initialize  $\phi_j^{(0)}$  (e.g., solve with no penalty)
  - Given  $\phi_j^{(k)}$ , we solve a **ridge problem** (here for  $J_2$  penalty)

$$\begin{aligned} \phi_j^{(k+1)} = & \arg \min_{\phi_j} \frac{1}{d_j^2} \sum_{i=1}^n \left( x_{ij} - \sum_{t=1}^{j-1} \phi_{j,t} x_{it} \right)^2 + \\ & + \lambda_1 \sum_{t=1}^{j-1} \frac{\phi_{j,t}^2}{|\phi_{j,t}^{(k)}|} + \lambda_2 \sum_{t=1}^{j-2} \frac{\phi_{j,t}^2}{|\phi_{j,t}^{(k)}| \cdot |\phi_{j,t+1}^{(k)}|} \end{aligned}$$

Tuning parameters selected on a **validation set**;  $\lambda_2$  is not important

## Simulation results

Three models (Huang et al., 2006):

$\Sigma_1$ : Identity

$\Sigma_2$ :  $d_j = 0.1$ ,  $\phi_{j,j-1} = 0.8$  and  $\phi_{jj'} = 0$  otherwise  
 $\Sigma_2^{-1}$  is tri-diagonal, corresponds to AR(1)

$\Sigma_3$ :  $d_j = 0.1$  and  $\phi_{jj'} = 0.5^{|j-j'|}$  (corresponds to MA(1))

- In order to compare to Lasso, use **quadratic loss**:

$$\Delta(\Sigma, \hat{\Sigma}) = \text{tr} \left( \Sigma^{-1} \hat{\Sigma} - I \right)^2$$

- $n = 100$ ,  $p = 30, 100$ , validation set size 100 (for selecting  $\lambda$ ), loss averaged over 50 replications

## Multivariate normal

	$p = 30$			$p = 100$		
	Sample	Lasso	AB	Sample	Lasso	AB
$\Sigma_1$	9.43(0.65)	0.55(0.12)	0.55(0.12)	100.8(3.4)	2.05(0.24)	2.08(0.23)
$\Sigma_2$	9.35(0.56)	1.89(0.23)	1.13(0.19)	103.6(2.1)	9.92(1.11)	4.02(0.28)
$\Sigma_3$	9.25(0.67)	8.93(1.06)	2.43(0.26)	103.3(3.4)	101.2(5.7)	8.87(0.58)

Multivariate  $t_3$ 

	$p = 30$			$p = 100$		
	Sample	Lasso	AB	Sample	Lasso	AB
$\Sigma_1$	28.3(19.1)	3.11(0.90)	3.24(0.87)	297.6(238.9)	12.2(4.5)	11.8(4.8)
$\Sigma_2$	40.7(30.0)	8.17(4.04)	6.19(3.01)	456.3(444.6)	46.3(38.1)	19.0(9.4)
$\Sigma_3$	25.8(19.2)	19.2(13.6)	6.48(2.42)	334.4(283.3)	144.8(113.0)	24.7(11.3)

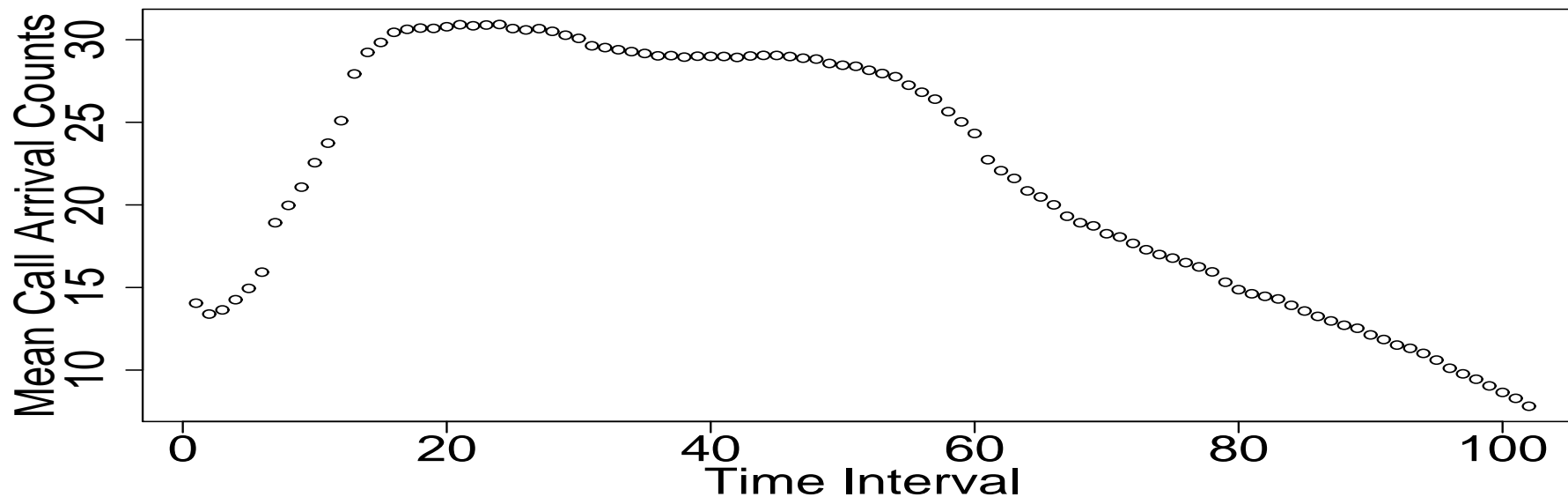
## Preserving sparsity

Percentage of true zeros in the estimator

	Cholesky factor $T$		$\Sigma^{-1}$	
	Lasso	AB	Lasso	AB
$\Sigma_1, p = 30$	99.8(0.3)	99.6(0.9)	99.8(0.3)	99.6(1.1)
$\Sigma_1, p = 100$	99.9(0.2)	100.0(0.1)	99.9(0.2)	100.0(0.1)
$\Sigma_2, p = 30$	71.7(4.3)	94.7(1.2)	35.5(10.2)	94.7(1.2)
$\Sigma_2, p = 100$	92.7(0.6)	99.0(0.7)	76.0(3.1)	99.0(0.7)

## Example: Call center data

- Collected from a call center in a U.S. financial organization for  $n = 239$  days in 2002 (Shen & Huang, 2005)
- Each day (7am - 12am) was divided into 102 time intervals
- $N_{ij}$ : number of calls during the  $j$ th time interval on the  $i$ th day;  
 $X_{ij} = \sqrt{N_{ij} + 1/4}$



**Goal:** forecast the call counts in the second half of the day using the arrival counts in the first half (Huang et al., 2006)

- Assume multivariate normality

$$\begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

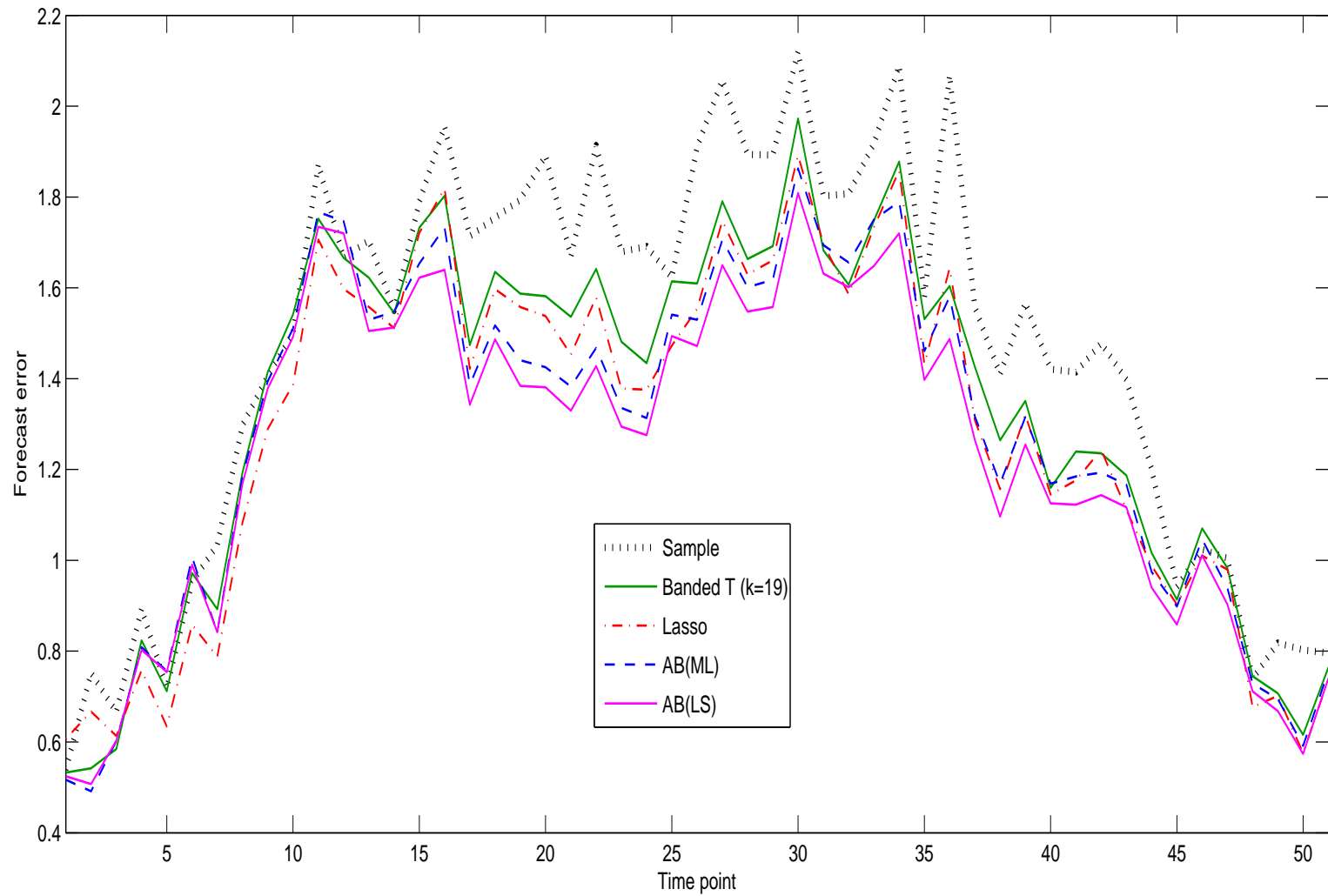
Then

$$E(\mathbf{x}^{(2)} | \mathbf{x}^{(1)}) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}^{(1)} - \mu_1)$$

- Divide the data into training set (January to October) and testing set (November and December):  $239 = 205 + 34$
- For each time interval  $j$ , the average absolute forecast error is

$$AE_j = \frac{1}{34} \sum_{i=206}^{239} |\hat{x}_{ij} - x_{ij}|$$

## Call center results



## Open problems and questions

- Analysis of **penalty** methods
- Estimators **invariant under permutations** of variables
- What **loss functions** are appropriate?
- **Direct optimization** approaches (semi-definite programming)
- What are the implications for PCA and other **applications**?
- Under what **conditions** would one benefit from using a particular estimator?