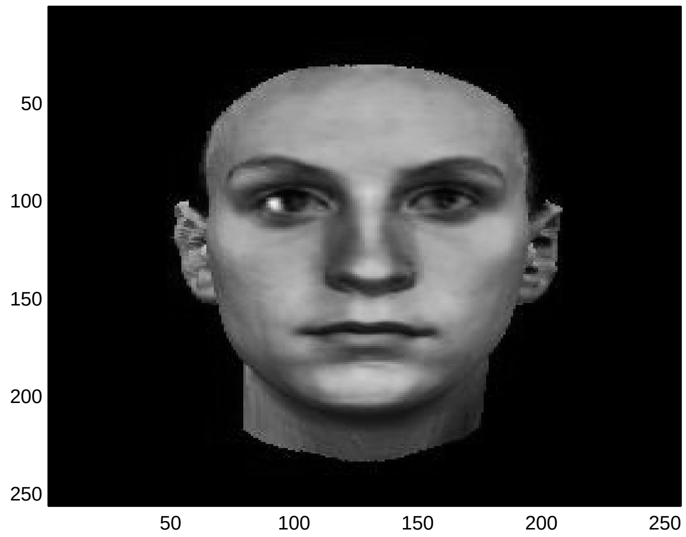
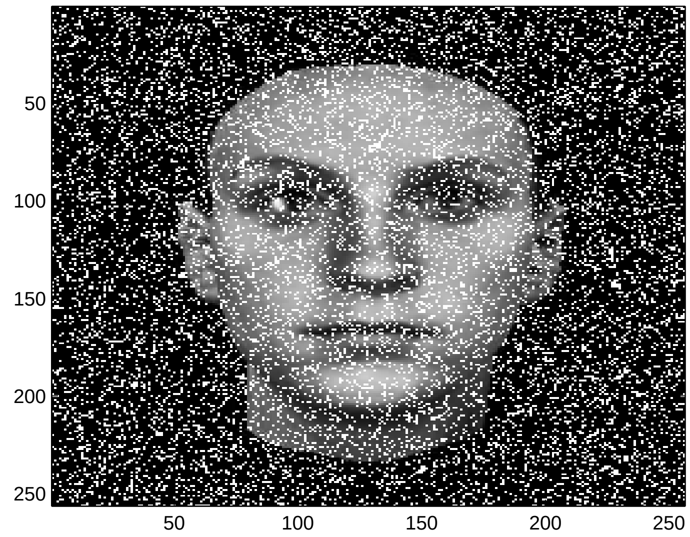
Image Denoising via Solution Paths





Joint work with Li Wang (PhD student, University of Michigan)

Image Denoising via Solution Paths

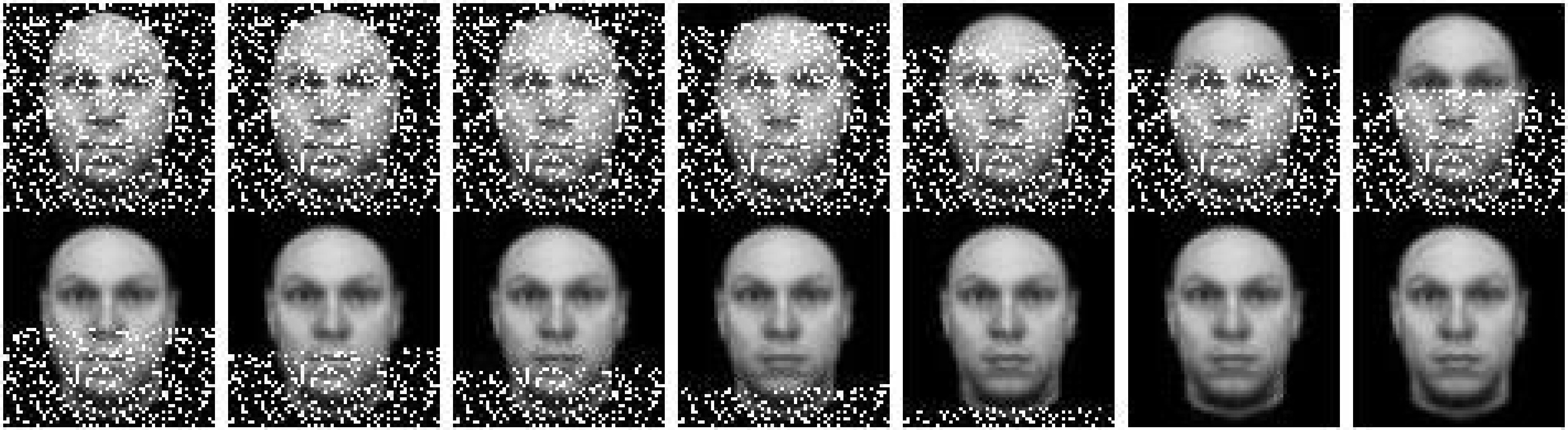


Image Denoising via Solution Paths

Sparse Covariance Estimation When Variables are Ordered

Ji Zhu
Assistant Professor
Statistics Department
University of Michigan

Acknowledgment

Joint work with Liza Levina (Assistant Professor, Department of Statistics, University of Michigan)

Outline

- Background
- Decomposing the covariance matrix
- Covariance estimation via regression
- LASSO penalty vs. AB penalty
- Numerical results
- Concluding remarks

Why Covariance Matrix

Many statistical and machine learning tools require an estimate of a covariance matrix.

- PCA
- LDA/QDA
- Graphical models
- $\vdots$

Basic Background

- Observe n i.i.d. samples, often from Normal$(\mathbf{0}, \mathbf{\Sigma}_{p\times p})$

$$\boldsymbol{X}^*_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{\Sigma}_{p\times p} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots \\ \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \cdots & \cdots \end{pmatrix}$$

Sample Covariance Matrix

- Wish to estimate $\boldsymbol{\Sigma}$
- Assuming the columns are centered, then the sample covariance matrix is

$$\hat{S} = \frac{1}{n-1} \boldsymbol{X}^{*\mathsf{T}} \boldsymbol{X}^{*}$$

Beyond the Sample Covariance Matrix

- Although the sample covariance matrix is unbiased, it can be extremely noisy, especially when p is large (Johnstone, 2001)
- Shrinkage methods
  - Haff, 1980; Dey & Srinivasan, 1985; Friedman, 1989; Ledoit & Wolf, 2003
  - Dempster, 1972; Pourahmadi, 1999; Wu & Pourahmadi, 2003; Meinshausen & Bühlmann, 2006; Huang et al., 2006

Our Focus

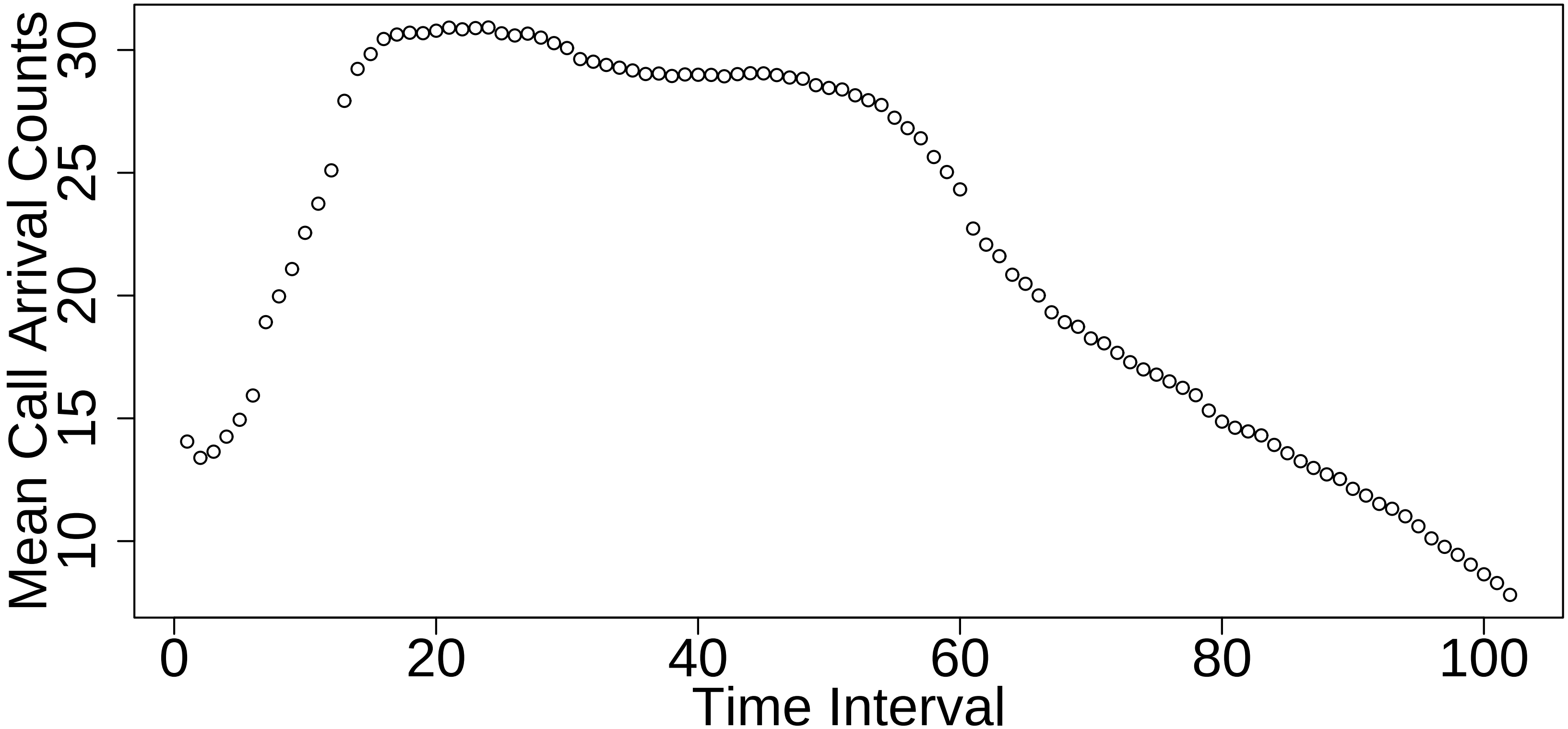
- We are interested in the case when there is an intrinsic order among the variables
- Examples:
  - $X_1, \ldots, X_p$ are information collected over time (longitudinal study)
  - The indices $1, \ldots, p$ represent a meaningful order (spectrum)

Example: Call Center

- Data were collected from a call center in a U.S. financial organization (Shen & Huang, 2005)
- Each day was divided into 102 time intervals
- x_{ij}: number of calls arrived during the jth time interval on the ith day
- $n = 239$ days in year 2002
- Forecast the call arrival counts in the later half of a day using the arrival counts in the early half of the day

Example: Call Center

Average mean counts over the 239 days



Example: Call Center

Assume multivariate normality

$$\begin{pmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

Then

$$\mathrm{E}(\boldsymbol{X}^{(2)} | \boldsymbol{X}^{(1)}) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}_1)$$

Example: Protein Mass Spectroscopy

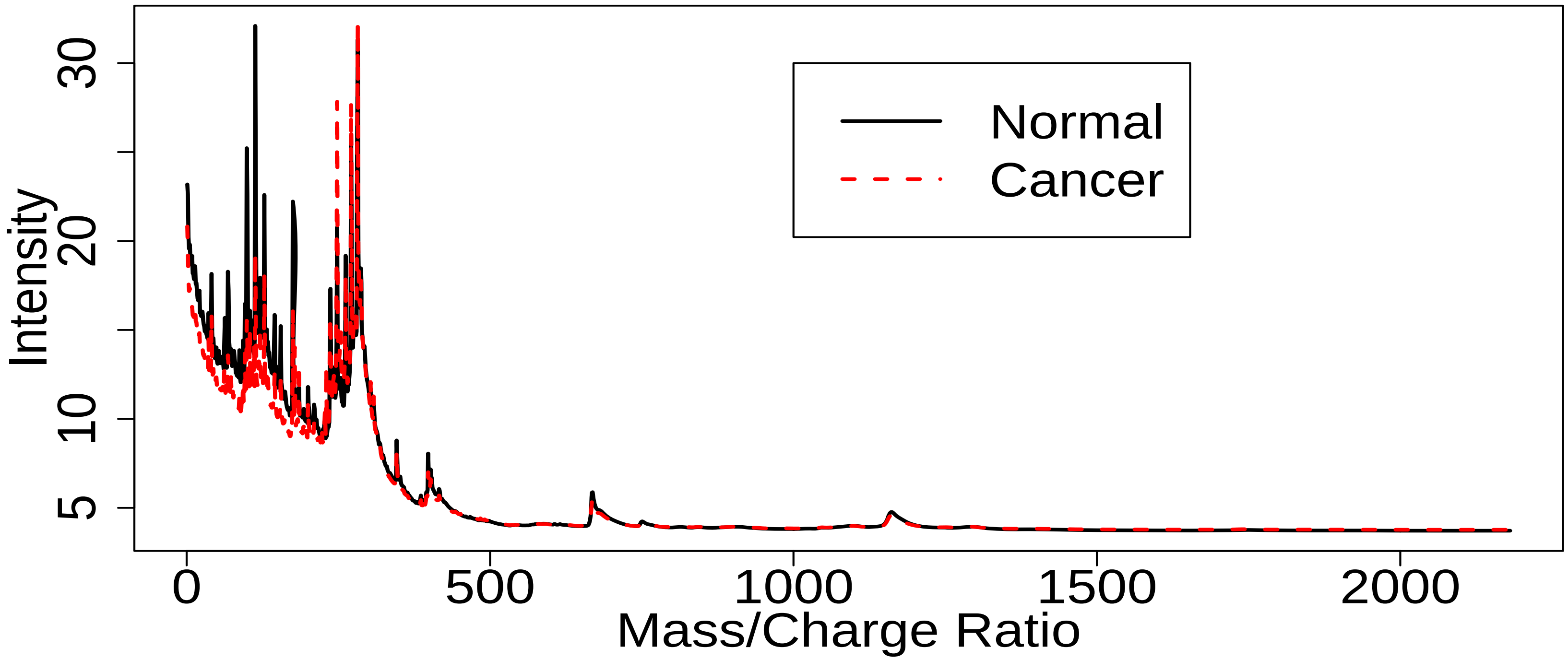
- Discriminate between healthy patients and prostate cancer patients using their blood serum samples
- For each sample i, x_{ij} is the intensity at the jth mass over charge ratio (m/z) of the constituent proteins
- Discriminant functions

$$\text{LDA} \quad : \quad \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k, \;\; k = 1, 2$$

$$\text{QDA} \quad : \quad \frac{1}{2} \ln |\boldsymbol{\Sigma}_k^{-1}| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k), \;\; k = 1, 2$$

Example: Protein Mass Spectroscopy

Average spectra profiles for healthy patients and those with prostate cancer



Outline

- ✔ Background
- Decomposing the covariance matrix
- Covariance estimation via regression
- LASSO penalty vs. AB penalty
- Numerical results
- Concluding remarks

Decomposition of Σ

The population covariance matrix can be re-written using the modified Cholesky decomposition (Pourahmadi 1999)

$$\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{D}$$

where $\boldsymbol{L}$ is a lower triangular matrix with diagonal elements equal to 1 and $\boldsymbol{D}$ is a diagonal matrix.

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & \cdots & \cdots \\ ? & 1 & 0 & \cdots \\ ? & ? & \ddots & 0 \\ ? & ? & \cdots & 1 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} d_1^2 & 0 & \cdots & \cdots \\ 0 & d_2^2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & d_p^2 \end{pmatrix}$$

Elements of $\boldsymbol{L}$ and $\boldsymbol{D}$

It turns out

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \phi_{21} X_1 + \epsilon_2 \\ X_3 &= \phi_{32} X_2 + \phi_{31} X_1 + \epsilon_3 \\ \vdots &\quad \vdots \\ X_p &= \phi_{p,p-1} X_{p-1} + \cdots + \phi_{p1} X_1 + \epsilon_p \end{aligned}$$

Hence $\boldsymbol{L}\boldsymbol{X} = \boldsymbol{\epsilon}$, where $L_{jj'} = -\phi_{jj'}$ for $j > j'$, and $d_j^2 = \mathrm{var}(\epsilon_j)$. Notice $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^\intercal \boldsymbol{D}^{-1} \boldsymbol{L}$.

Outline

- ✔ Background
- ✔ Decomposing the covariance matrix
- Covariance estimation via regression
- LASSO penalty vs. AB penalty
- Numerical results
- Concluding remarks

Negative Log-likelihood

Denote $\boldsymbol{x} = (x_1, \ldots, x_p)^\mathsf{T}$. If we assume Normality, the negative log-likelihood can be written as

$$\begin{aligned}\ell(\boldsymbol{x}, \boldsymbol{\Sigma}) &= \ln|\boldsymbol{\Sigma}| + \boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \\ &= \ln|\boldsymbol{D}| + \boldsymbol{x}^\mathsf{T}\boldsymbol{L}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{x} \\ &= \sum_{j=1}^{p} \ln d_j^2 + \sum_{j=1}^{p} \frac{\epsilon_j^2}{d_j^2}\end{aligned}$$

Estimating $\phi_{jj'}$ and d_j^2

The negative log-likelihood on the training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

$$\begin{aligned} \ell(\boldsymbol{X}^*, \boldsymbol{\Sigma}) &= n \sum_{j=1}^{p} \ln d_j^2 + \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\epsilon_{ij}^2}{d_j^2} \\ &= \sum_{j=1}^{p} \ell_j(\boldsymbol{X}^*, \boldsymbol{\Sigma}) \end{aligned}$$

Estimating $\phi_{jj'}$ and d_j^2

The negative log-likelihood decomposes into

$$j = 1 \qquad \ell_1(\boldsymbol{X}^*, \boldsymbol{\Sigma}) = nd_1^2 + \frac{1}{d_1^2}\sum_{i=1}^{n} x_{i1}^2$$

$$j > 1 \qquad \ell_j(\boldsymbol{X}^*, \boldsymbol{\Sigma}) = nd_j^2 + \frac{1}{d_j^2}\sum_{i=1}^{n}(x_{ij} - \phi_{j,j-1}x_{i,j-1}\cdots - \phi_{j1}x_{i1})^2$$

We can minimize them separately to find $\hat{\phi}_{jj'}$ and $\hat{d}_j^2$ (Each is essentially an OLS).

Outline

- ✔ Background
- ✔ Decomposing the covariance matrix
- ✔ Covariance estimation via regression
- LASSO penalty vs. AB penalty
- Numerical results
- Concluding remarks

Regularization

In practice, shrinkage is necessary, so we consider

$$\min_{\boldsymbol{\phi}_j, d_j} \ell_j(\boldsymbol{X}^*, \boldsymbol{\Sigma}) + \lambda \cdot J(\boldsymbol{\phi}_j)$$

where $J(\boldsymbol{\phi}_j)$ is a penalty term.

LASSO Penalty

Using the L_1-norm (LASSO) penalty (Huang et al. 2006)

$$J(\boldsymbol{\phi}_j) = \sum_{j'=1}^{j-1} |\phi_{jj'}|$$

- Shrinkage
- Sparseness: some $\hat{\phi}_{jj'} = 0$
- Sparse in $\boldsymbol{L}$, not necessarily in $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^\top \boldsymbol{D}^{-1} \boldsymbol{L}$

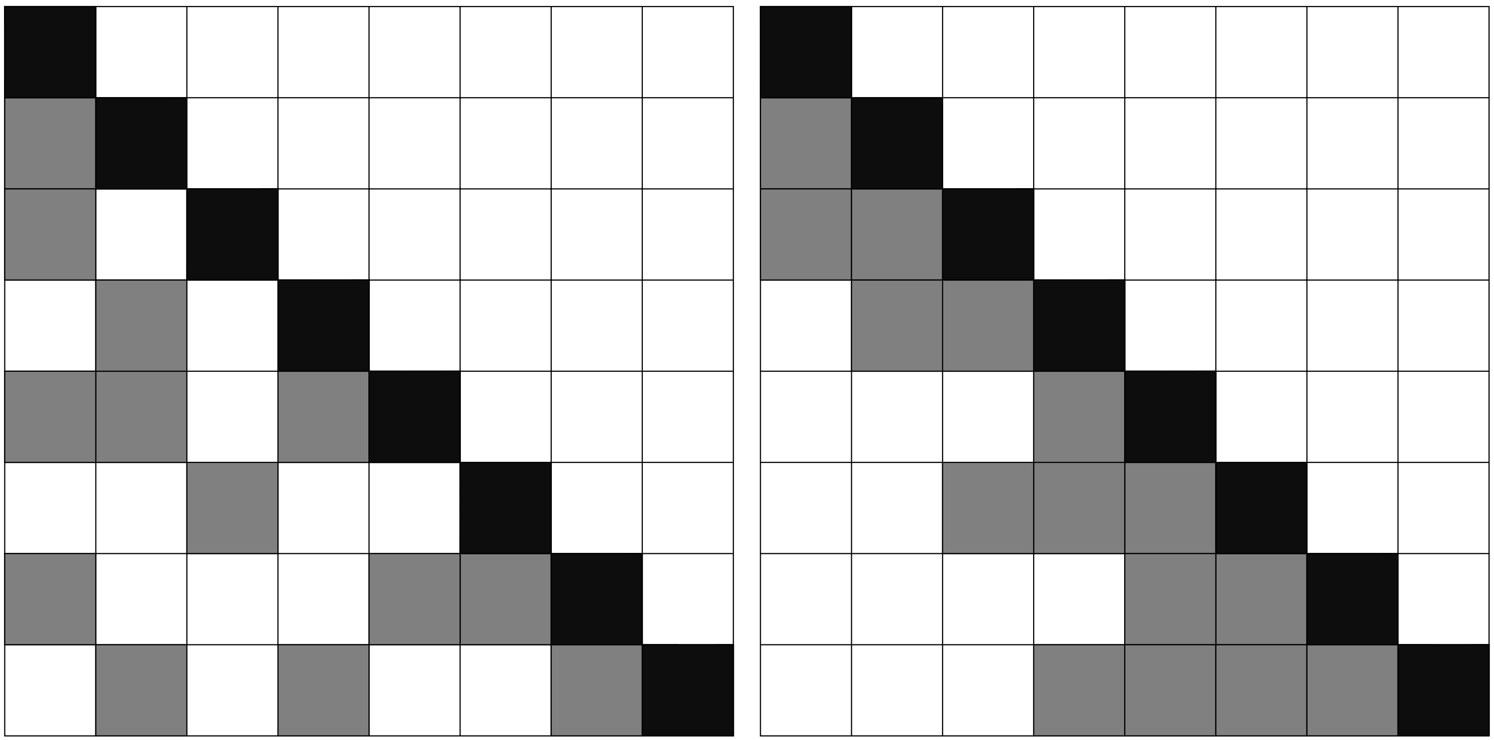
Incorporating the Order Information

We propose (Hierarchical LASSO)

$$J(\boldsymbol{\phi}_j) = |\phi_{j,j-1}| + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + \cdots + \frac{|\phi_{j,1}|}{|\phi_{j,2}|}$$

- Shrinkage
- Sparseness: If $\hat{\phi}_{jk} = 0$ for some k, then $\hat{\phi}_{jj'} = 0$ for all $j' < k$.
- Sparse in $\boldsymbol{L}$ and $\boldsymbol{\Sigma}^{-1}$
- Hierarchical LASSO $\Longrightarrow$ Adaptive Banding of $\boldsymbol{\Sigma}^{-1}$

LASSO vs Adaptive Banding (AB)



Estimation

The penalized negative log-likelihood on $\ell_j(\boldsymbol{X}^*, \boldsymbol{\Sigma})$ becomes

$$\min_{\boldsymbol{\phi}_j, d_j} \quad n \ln d_j^2 + \frac{1}{d_j^2} \sum_{i=1}^{n} (x_{ij} - \phi_{j,j-1} x_{i,j-1} \cdots - \phi_{j1} x_{i1})^2 + \\ + \lambda \cdot \left(|\phi_{j,j-1}| + \sum_{j'=2}^{j-1} \frac{|\phi_{j,j'-1}|}{|\phi_{j,j'}|} \right)$$

Iterative Procedure

1. Fix $\boldsymbol{\phi}_j$, solve for d_j
2. Fix d_j, solve for $\boldsymbol{\phi}_j$
3. Iterate between 1 and 2 until convergence

Step 1

When $\boldsymbol{\phi}_j$ is fixed, $n\hat{d}_j^2$ is the residual sum of squares

$$\hat{d}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \phi_{j,j-1}x_{i,j-1}\cdots - \phi_{j1}x_{i1})^2$$

Step 2

When d_j is fixed, the criterion becomes

$$\min_{\boldsymbol{\phi}_j} \quad \frac{1}{d_{jj}^2}\sum_{i=1}^{n}(x_{ij} - \phi_{j,j-1}x_{i,j-1}\cdots - \phi_{j1}x_{i1})^2 + \\ +\lambda\cdot\left(|\phi_{j,j-1}| + \sum_{j'=2}^{j-1}\frac{|\phi_{j,j'-1}|}{|\phi_{j,j'}|}\right)$$

Need another Iterative Procedure

1. Initialize $\boldsymbol{\phi}_j^{(0)}$

2. Given $\boldsymbol{\phi}_j^{(k)}$, we solve (a ridge problem)

$$\begin{aligned}\boldsymbol{\phi}_j^{(k+1)} &= \arg\min_{\boldsymbol{\phi}_j} \frac{1}{d_j^2}\sum_{i=1}^{n}(x_{ij} - \phi_{j,j-1}x_{i,j-1}\cdots - \phi_{j1}x_{i1})^2 + \\ &\quad +\lambda\cdot\left(\frac{\phi_{j,j-1}^2}{|\phi_{j,j-1}^{(k)}|} + \sum_{j'=2}^{j-1}\frac{\phi_{j,j'-1}^2}{|\phi_{j,j'-1}^{(k)}|\cdot|\phi_{j,j'}^{(k)}|}\right)\end{aligned}$$

3. $k \leftarrow k+1$ and go to 2 until convergence

Similar to Fan & Li (2001)

Regression Variant

Instead of using the negative log-likelihood, we can fit a penalized regression model directly (without d_j)

$$\begin{aligned}\hat{\boldsymbol{\phi}}_j &= \arg\min_{\boldsymbol{\phi}_j} \sum_{i=1}^{n}(x_{ij} - \phi_{j,j-1}x_{i,j-1} \cdots - \phi_{j1}x_{i1})^2 + \\ &\quad +\lambda \cdot \left(|\phi_{j,j-1}| + \sum_{j'=2}^{j-1} \frac{|\phi_{j,j'-1}|}{|\phi_{j,j'}|} \right)\end{aligned}$$

and set

$$\hat{d}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \hat{\phi}_{j,j-1}x_{i,j-1} \cdots - \hat{\phi}_{j1}x_{i1})^2$$

Outline

- ✔ Background
- ✔ Decomposing the covariance matrix
- ✔ Covariance estimation via regression
- ✔ LASSO penalty vs. AB penalty
- Numerical results
- Concluding remarks

Simulation Setup

$\mathbf{\Sigma}_1$ Identity

$\mathbf{\Sigma}_2$ $d_j = 0.01$, $\phi_{j,j-1} = 0.8$ and $\phi_{jj'} = 0$ otherwise (so $\mathbf{\Sigma}_2^{-1}$ is tri-diagonal)

$\mathbf{\Sigma}_3$ $d_j = 0.01$ and $\phi_{jj'} = 0.5^{|j-j'|}$ (non-sparse)

Simulation Setup

- $n = 100$ training data, 100 validation data
- $p = 30$ and 100
- $\boldsymbol{X} \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma})$ and Multivariate T_3
- Compare sample covariance, LASSO, AB-Lik (likelihood-based) and AB-Reg (regression-based)
- Entropy loss (Anderson 2003)

$$\Delta_E(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = \text{tr}\left(\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}\right) - \ln\left|\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}\right| - p$$

Simulation Result: Σ_1

p	Sample	LASSO	AB-Lik	AB-Reg
	Multivariate Normal			
30	5.24(0.30)	0.30(0.08)	0.31(0.08)	0.31(0.08)
100	133.8(3.2)	0.99(0.16)	0.99(0.16)	0.99(0.16)
	Multivariate T_3			
30	15.4(5.8)	2.72(2.69)	2.80(3.95)	2.46(2.55)
100	206.8(27.3)	9.64(7.45)	9.22(6.79)	9.08(6.60)

Simulation Result: Σ_2

p	Sample	LASSO	AB-Lik	AB-Reg
	Multivariate Normal			
30	5.32(0.33)	1.10(0.19)	0.69(0.15)	0.72(0.14)
100	133.0(3.2)	5.43(0.52)	2.29(0.24)	2.46(0.24)
	Multivariate T_3			
30	18.7(16.4)	7.7(14.1)	7.6(16.4)	5.8(10.6)
100	208.5(45.3)	26.1(27.3)	14.9(24.2)	14.1(16.6)

Simulation Result: Σ_3

p	Sample	LASSO	AB-Lik	AB-Reg
	Multivariate Normal			
30	5.14(0.31)	3.16(0.34)	1.20(0.12)	1.18(0.13)
100	133.5(3.4)	24.7(3.5)	4.33(0.25)	4.25(0.24)
	Multivariate T_3			
30	15.0(5.6)	14.4(5.3)	5.00(3.50)	4.69(2.72)
100	200.8(26.1)	164.7(22.4)	18.5(10.9)	17.3(9.5)

Percentage of Zeros (Normal)

(# zeros in the estimate / # zeros in the truth)

	$\boldsymbol{\Sigma}^{-1}$	
p	LASSO	AB
$\boldsymbol{\Sigma}_1, 30$	99.9(0.2)%	99.4(1.1)%
$\boldsymbol{\Sigma}_1, 100$	99.9(0.1)%	99.9(0.1)%
$\boldsymbol{\Sigma}_2, 30$	31.9(7.2)%	95.1(1.4)%
$\boldsymbol{\Sigma}_2, 100$	76.4(3.6)%	98.9(0.2)%

Percentage of Zeros (T_3)

(# zeros in the estimate / # zeros in the truth)

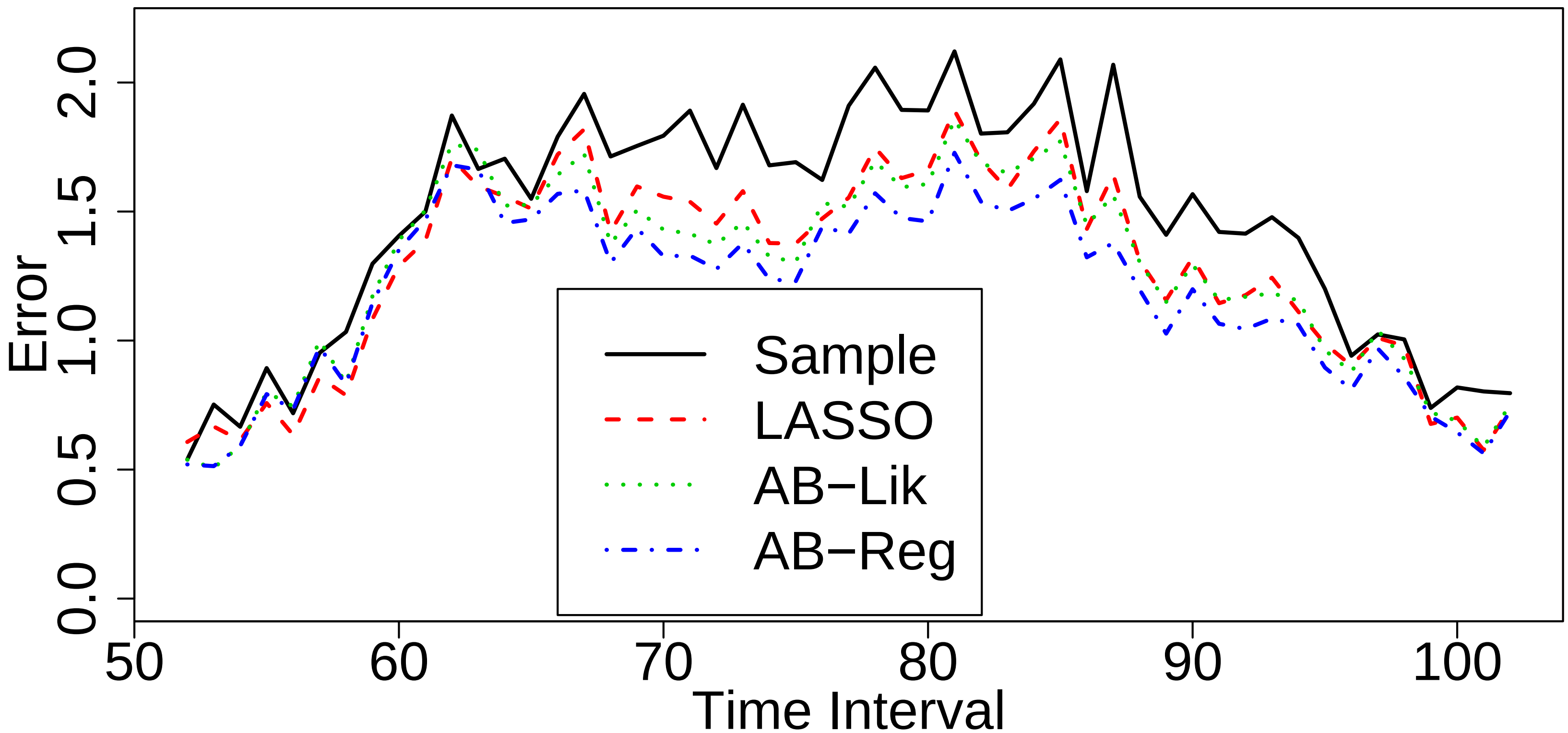
	$\mathbf{\Sigma}^{-1}$	
p	LASSO	AB
$\mathbf{\Sigma}_1, 30$	93.8(11.3)%	98.0(5.9)%
$\mathbf{\Sigma}_1, 100$	99.3(2.1)%	99.9(0.2)%
$\mathbf{\Sigma}_2, 30$	45.3(13.5)%	94.9(4.0)%
$\mathbf{\Sigma}_2, 100$	37.4(4.5)%	98.6(1.0)%

Call Center Data

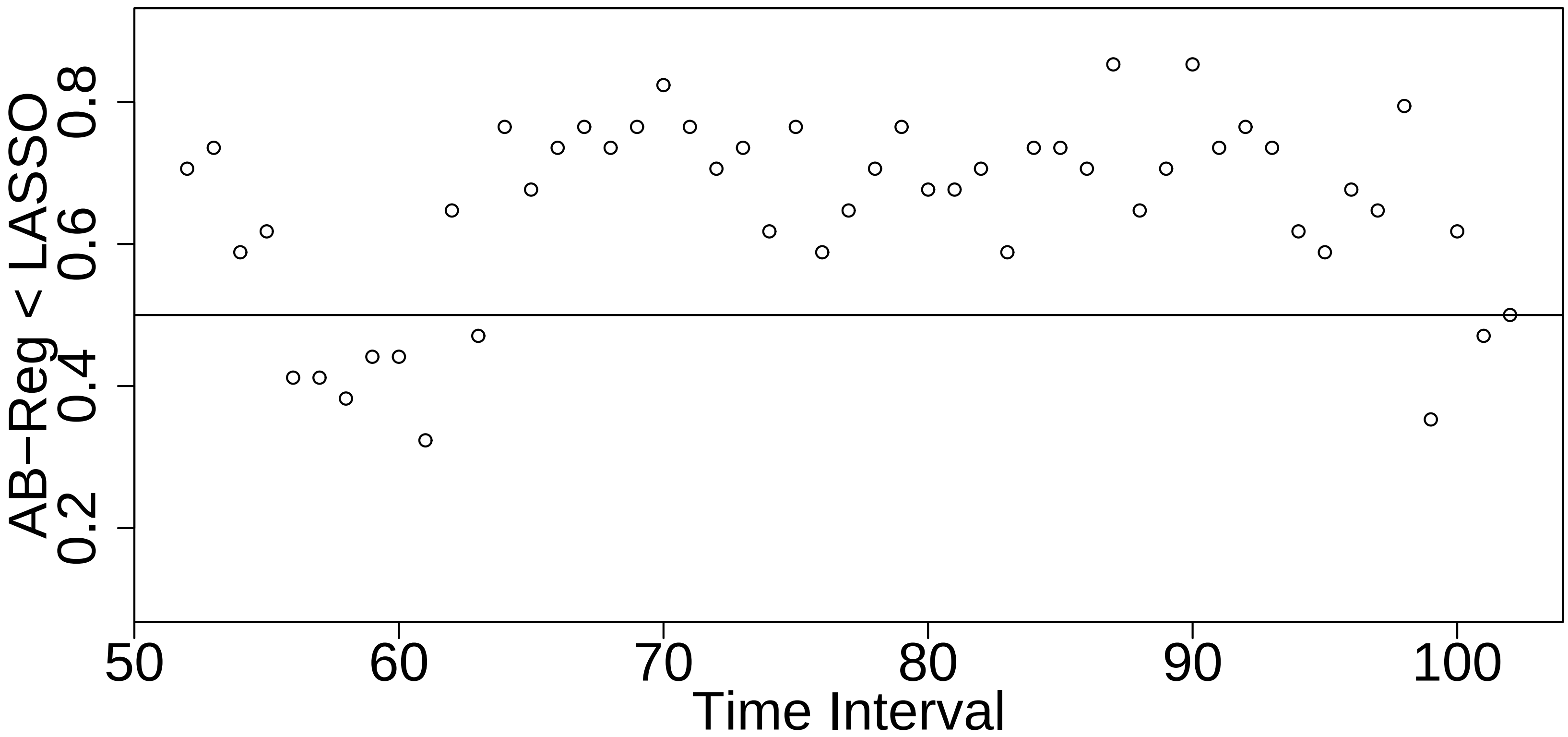
- Divide the data into training set (January to October) and testing set (November and December): $239 = 205 + 34$
- Use the 51 arrival counts in the early half of a day to forecast the 51 arrival counts in the later half of the day
- For each time interval j, the forecast error is

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

Call Center Data



Call Center Data



Concluding Remarks

- Adaptive Banding: achieves shrinkage, incorporates order information, preserves sparsity; better than LASSO
- Open issues
  - Evaluating the performance (other loss functions)
  - Multiple λ_j's
  - Imposing other structures