Adaptive Estimation of Large Covariance Matrices

Ming Yuan

School of Industrial and Systems Engineering

Georgia Institute of Technology

myuan@isye.gatech.edu

http://www.isye.gatech.edu/~myuan

High Dimensional Covariance Matrix Estimation

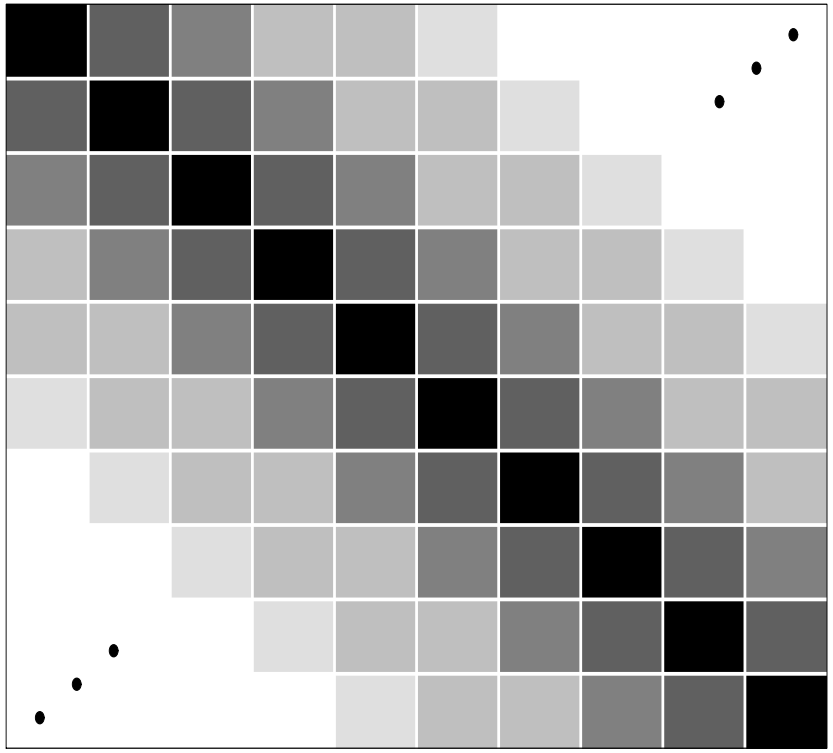
- Covariance Matrix Estimation
  - Data – a sample of n independent copies of a r.v. $X \in \mathbb{R}^d$
  - Covariance matrix – $\Sigma = \mathrm{cov}(X)$
  - Sample covariance matrix – $\bar{\Sigma}^{\mathrm{Sample}}$
- High dimensionality
  - Classical paradigm – d fixed and $n \longrightarrow \infty$
  - High dimensional problems – $d \gtrsim n$
- Parameter reduction through sparsity
  - Sparsity in Cholesky factors
  - Sparsity in inverse covariance matrix
  - Sparsity in covariance matrix
    - Bandable covariance matrix

OUTLINE

- Why adaptive estimation?
- How to achieve adaption?
  - Block thresholding scheme
  - High level ideas of the proof
- What else can we do?
  - Optimality of banding
  - Beyond normality

Bandable Covariance Matrices

BANDABLE COVARIANCE MATRICES



- Bounded eigenvalues

$$M_0^{-1} \leq \lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma) \leq M_0$$

- Decaying off-diagonal entries

$$\max_j \sum_i \{|\sigma_{ij}| : |i-j| \geq k\} \leq M k^{-\alpha}$$

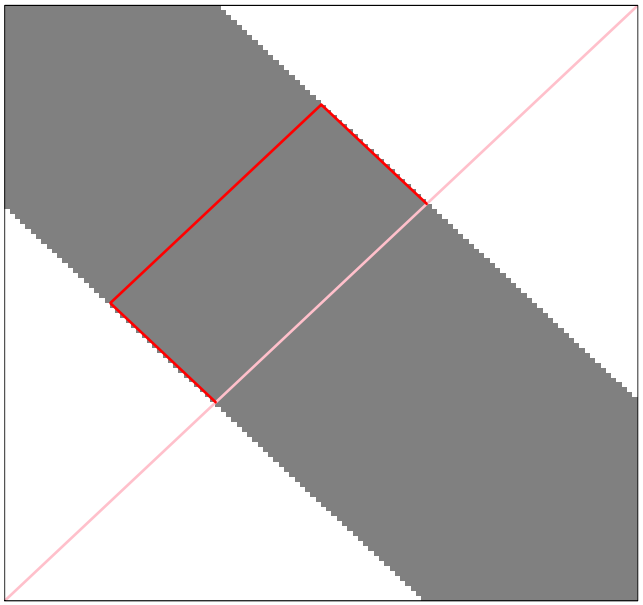
- Alternative specification

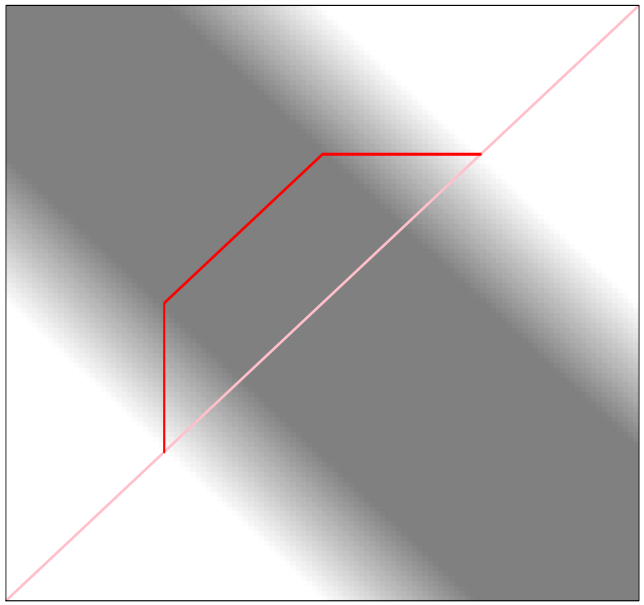
$$|\sigma_{ij}| \leq M|i-j|^{-(\alpha+1)} \qquad \forall i \neq j$$

(Bickel and Levina, 2008)

Estimation Methods

$$\hat{\Sigma} = \bar{\Sigma}^{\text{Sample}} \circ W$$





- Banding (Bickel and Levina, 2008)

$$W = B_k := (\mathbb{I}(|i-j| \le k))$$

- Tapering (Cai, Zhang and Zhou, 2010)

$$W = T_k := (\frac{2}{k}\{(k-|i-j|)_+ - (\frac{k}{2} - |i-j|)_+\})$$

Minimax Optimal Rate of Convergence

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \asymp \min\left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log d}{n}, \frac{d}{n} \right\},$$

(Cai, Zhang and Zhou, 2010)

- Tapering

$$k \asymp n^{1/(2\alpha+1)} \Longrightarrow \text{Minimax optimality}$$

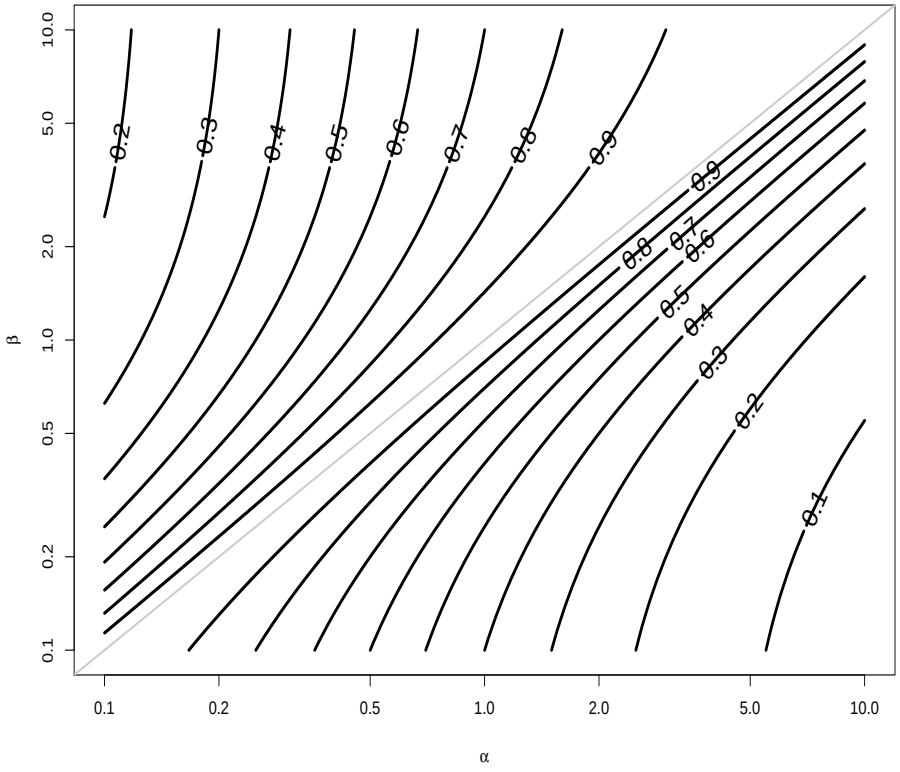
- Banding
  - Original choice by Bickel and Levina (2008)

$$k \asymp (n/\log d)^{1/(2(\alpha+1))} \Longrightarrow \|\hat{\Sigma} - \Sigma\| \lesssim_p \left(\frac{\log d}{n}\right)^{\alpha/(2\alpha+2)}$$

  - Optimality as byproduct of our analysis

$$k \asymp n^{1/(2\alpha+1)} \Longrightarrow \text{Minimax optimality}$$

IMPORTANCE OF KNOWING α



- True decay rate – α
- Tuning parameter – $k_\beta \asymp n^{1/(2\beta+1)}$
- Relative efficiency in rate of convergence
  - Optimal choice

$$-\frac{\log \sup_{\Sigma\in\mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma}_{k_\alpha} - \Sigma\|^2}{\log n} = \frac{2\alpha}{2\alpha+1}$$

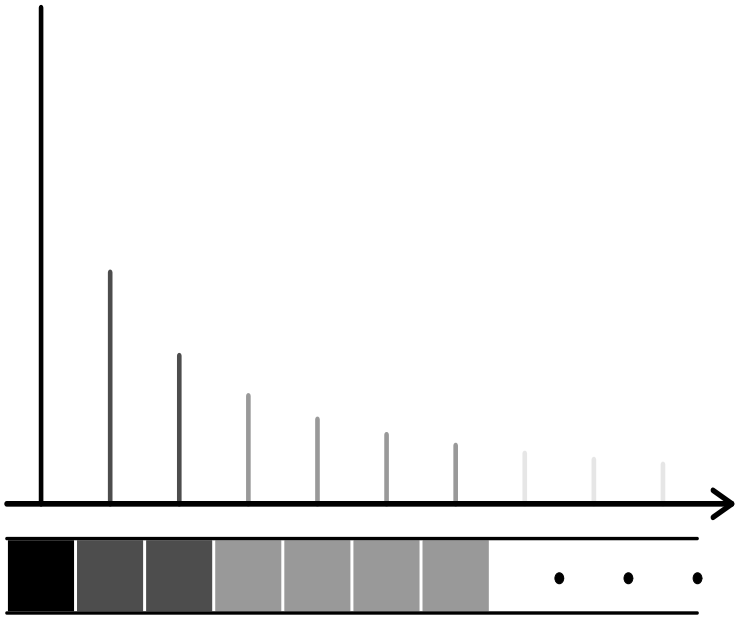
  - Suboptimal choice

$$\underbrace{-\frac{\log \sup_{\Sigma\in\mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma}_{k_\beta} - \Sigma\|^2}{\log n}}_{\text{achieved rate}} \Big/ \underbrace{\frac{2\alpha}{2\alpha+1}}_{\text{optimal rate}}$$

Adaptive estimation – A single estimator to achieve the optimal rate of convergence over all α

BLOCK THRESHOLDING

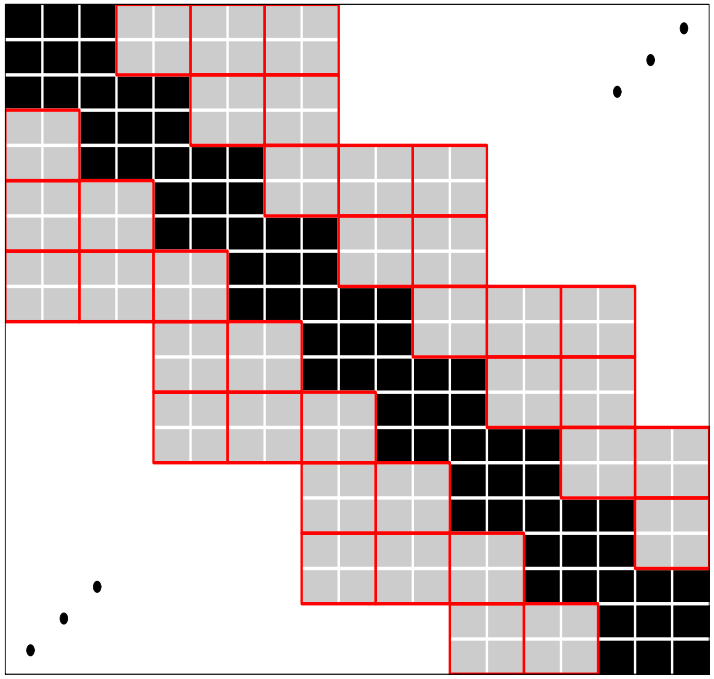
BLOCK THRESHOLDING IN FUNCTION ESTIMATION



- Smooth functions – Fourier coefficient $f_k \asymp k^{-2\alpha}$
- Block thresholding
  - Creating blocking by doubling the block sizes
  - Block-by-block thresholding
- Overall error can be decomposed into blocks
- See, e.g., Efromovich (1985)

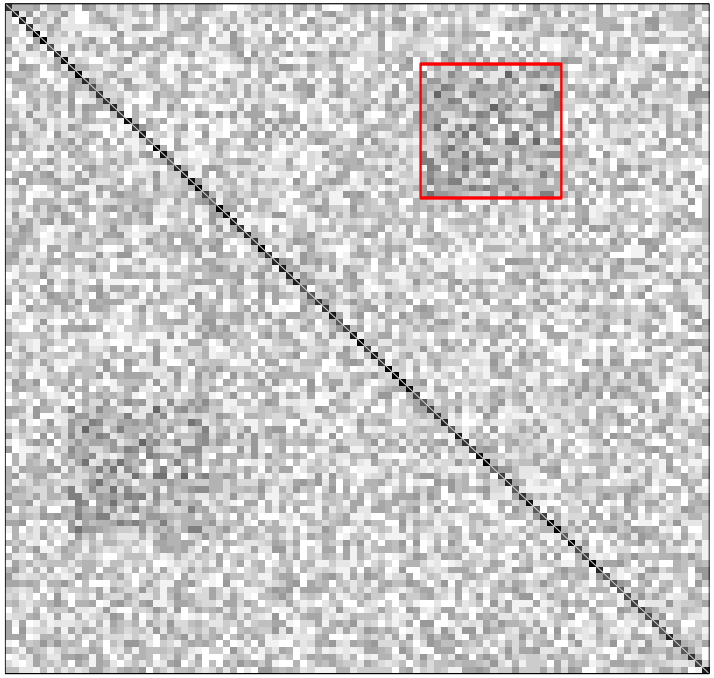
How to do it for covariance matrix – blocking and thresholding?

BLOCKING SCHEME



- Start by constructing blocks of size $k \times k$ where $k = \log d$
  - Create blocks on the diagonal
  - Create more blocks successively
    - Two or one in an alternating fashion
- Double block size $k = 2 \log d$ and create blocks successively
  - Three or two in an alternating fashion
- Continue until the whole matrix is covered

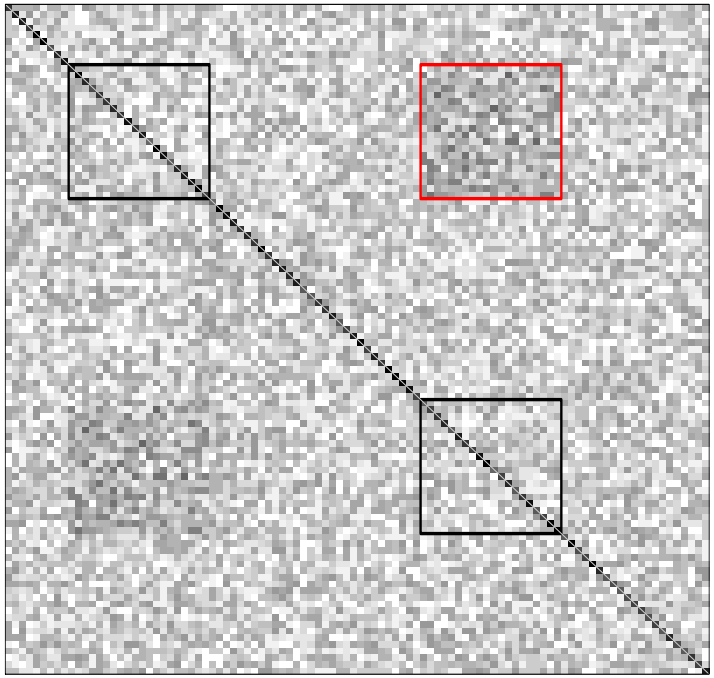
BLOCK THRESHOLDING



- For each block B
  - Kill the block – $\hat{\Sigma}_B = \mathbf{0}$
  - Keep the block – $\hat{\Sigma}_B = \bar{\Sigma}_B^{\text{Sample}}$
- Ideally, keep blocks if Σ_B is large
  - Large in which sense
  - Do not observe Σ_B directly
  - Large in size $\rightsquigarrow$ large in magnitude
- Deemed large if

$$\|\bar{\Sigma}_B^{\text{Sample}}\| \geq \lambda \left(\frac{s(B) + \log d}{n} \right)^{1/2}$$

SCALE-FREE BLOCK THRESHOLDING



- λ may depend on M or M_0
- Measure the magnitude of Σ_B relative to the corresponding diagonal blocks
  - Let $B = I \times J$ – Diagonal blocks $\Sigma_{I\times I}$ and $\Sigma_{J\times J}$

$$\lambda = \lambda_0 \left(\|\bar{\Sigma}^{\text{Sample}}_{I\times I}\| \|\bar{\Sigma}^{\text{Sample}}_{J\times J}\| \right)^{1/2}$$

  - $\lambda_0 > 5.44$
- Alternatively thresholding could be based on sample canonical correlation between X_I and X_J

ADAPTIVITY

ADAPTIVITY

MAIN RESULT Let $\hat{\Sigma}$ be the block thresholding estimator of Σ. Then

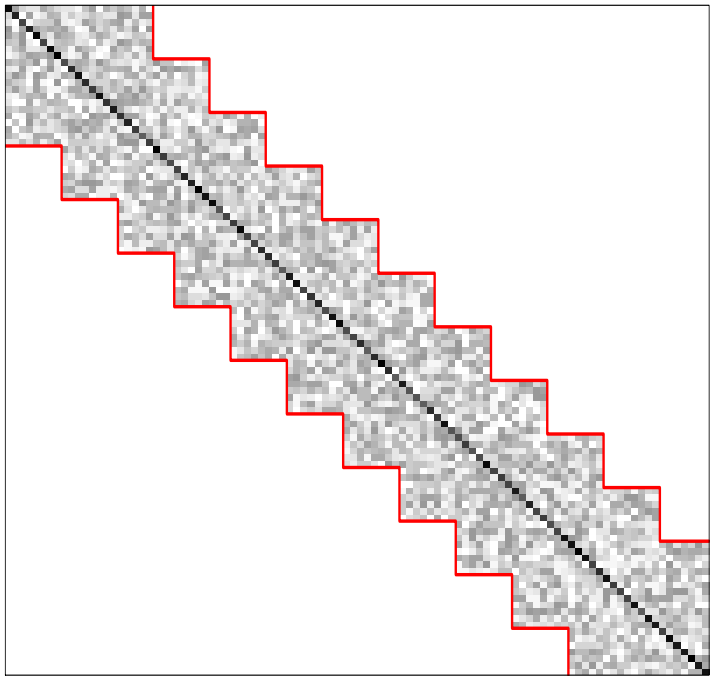
$$\sup_{\Sigma \in \mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \lesssim \min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log d}{n}, \frac{d}{n}\right\},$$

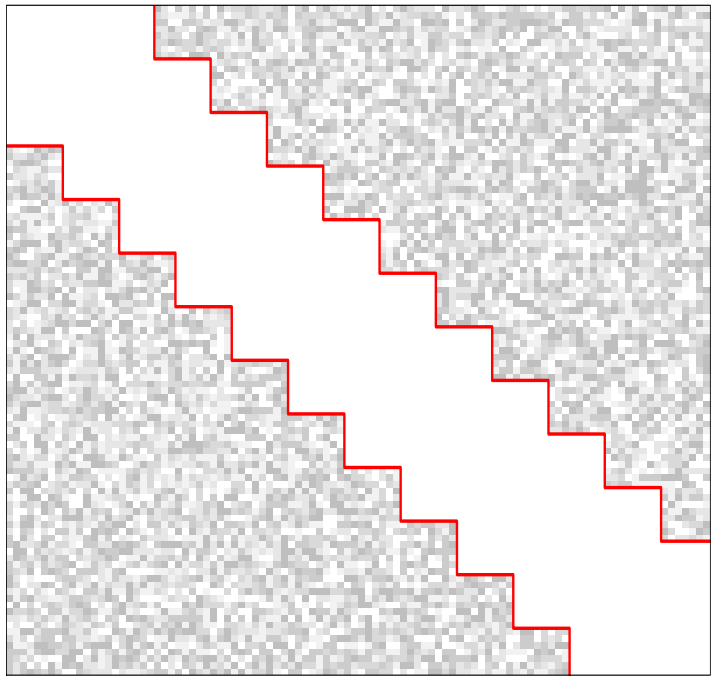
for all $\alpha > 0$.

- Same rate holds for estimating inverse covariance matrix
- Difficulty – $\|\hat{\Sigma} - \Sigma\|^2$ cannot be decomposed into blocks
- Solution – treat blocks of the same size in unison

$$S(\Delta; l) = \sum_{B \in \mathcal{B}: s(B) = 2^{l-1} \log d} \Delta_B$$

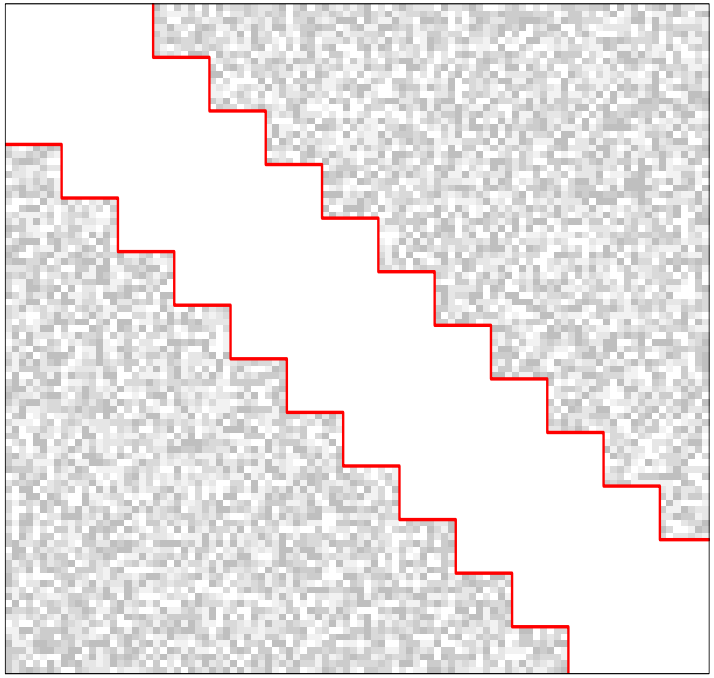
SMALL BLOCKS VS LARGE BLOCKS





$$\|\hat{\Sigma}-\Sigma\| = \left\|\sum_l S(\hat{\Sigma}-\Sigma,l)\right\| \le \underbrace{\left\|\sum_{l\le L} S(\hat{\Sigma}-\Sigma,l)\right\|}_{\text{small blocks}} + \underbrace{\left\|\sum_{l>L} S(\hat{\Sigma}-\Sigma,l)\right\|}_{\text{large blocks}}$$

LARGE BLOCKS



- Large blocks are necessarily far away from diagonal

$$\min_{(i,j)\in B} |i-j| \geq s(B)$$

- As a result

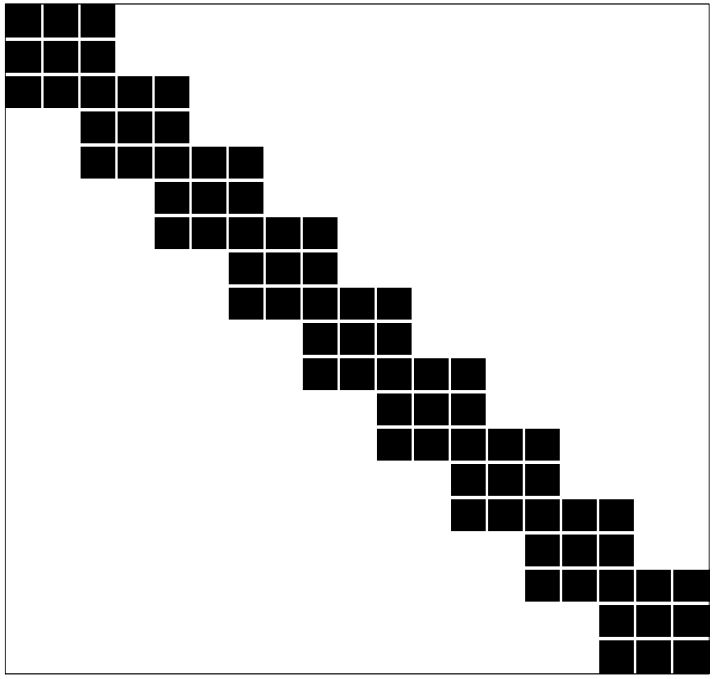
$$\|\Sigma_B\| \leq M s(B)^{-\alpha}$$

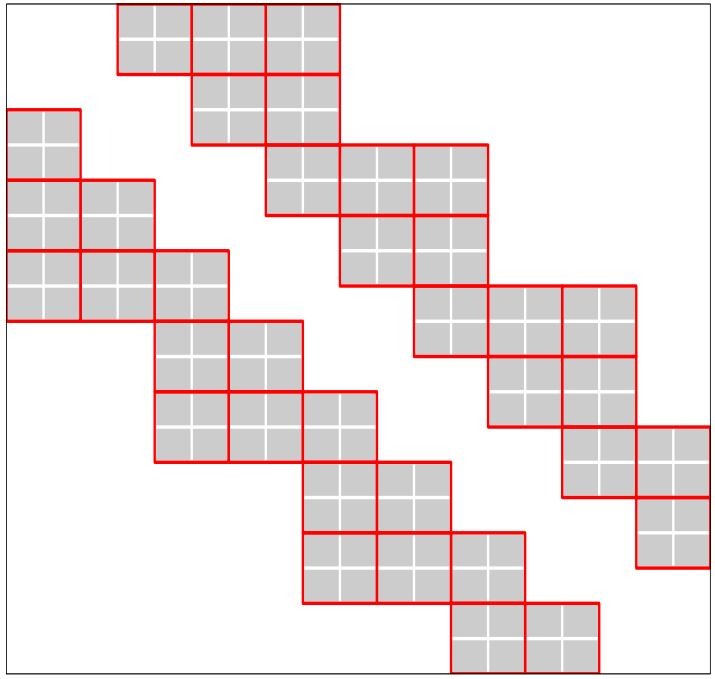
- Concentration inequality

$$\frac{\|\bar{\Sigma}_B^{\text{Sample}} - \Sigma_B\|^2}{\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|} \lesssim_{\text{w.h.p.}} \frac{s(B)+\log d}{n}$$

$$\Longrightarrow \left\|\sum_{l>L} S(\hat{\Sigma}-\Sigma,l)\right\| =_{\text{w.h.p.}} 0 \Rightarrow \text{bounds for } \mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma}-\Sigma,l)\right\|^2$$

SMALL BLOCKS





$$\left\| \sum_{l \le L} S(\hat{\Sigma} - \Sigma, l) \right\| \le \sum_{l \le L} \left\| S(\hat{\Sigma} - \Sigma, l) \right\|$$

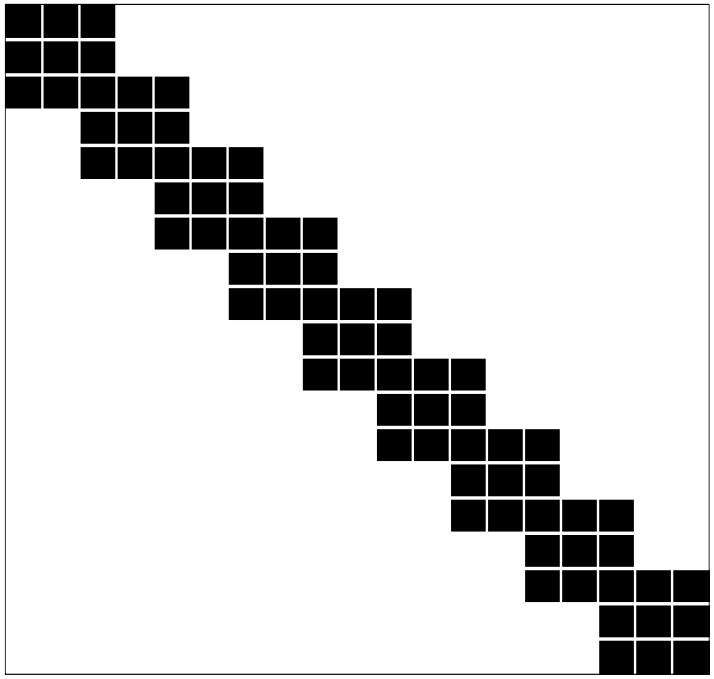
Norm Compression Inequality

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1G} \\ A_{21} & A_{22} & \dots & A_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ A_{G1} & A_{G2} & \dots & A_{GG} \end{pmatrix} \mapsto \mathcal{N}(A) = \begin{pmatrix} \|A_{11}\| & \|A_{12}\| & \dots & \|A_{1G}\| \\ \|A_{21}\| & \|A_{22}\| & \dots & \|A_{2G}\| \\ \vdots & \vdots & \ddots & \vdots \\ \|A_{G1}\| & \|A_{G2}\| & \dots & \|A_{GG}\| \end{pmatrix}$$

where $A_{jk} \in \mathbb{R}^{d_j \times d_k}$. Then

$$\|A\| \le \|\mathcal{N}(A; d_1, \dots, d_G)\|$$

Bounding $S(\hat{\Sigma} - \Sigma, 1)$

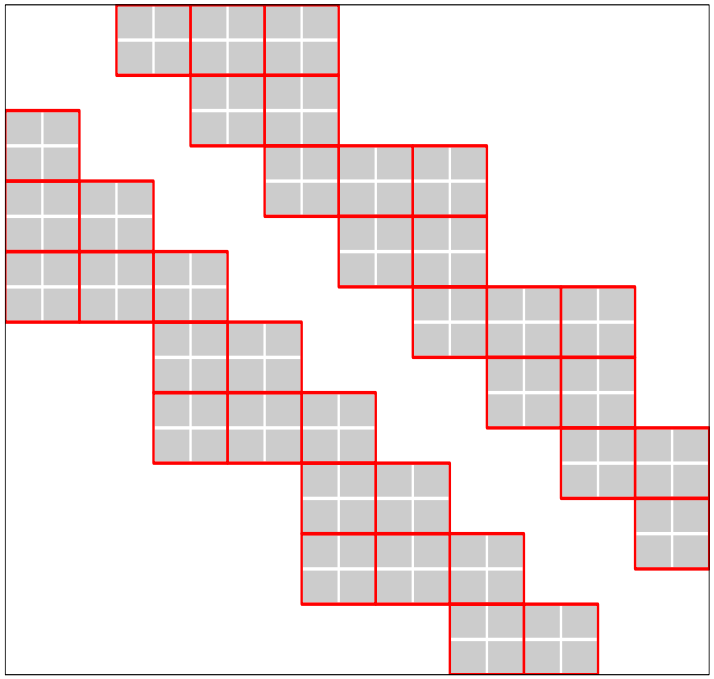


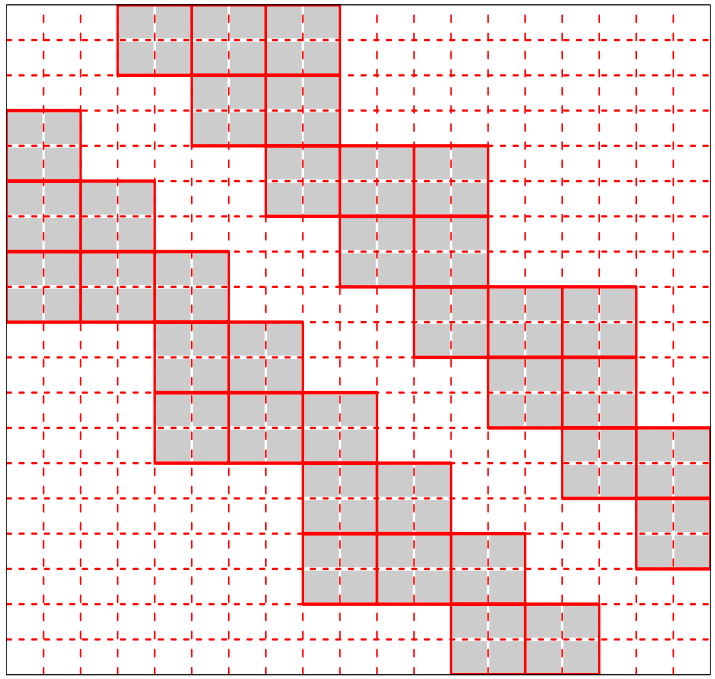
- Take $d_1, \ldots, d_G = \log d$
- $\mathcal{N}(S)$ is a 3-band matrix

$$\|\mathcal{N}(S)\| \leq 5 \max_{j,k} |\mathcal{N}(S)_{jk}|$$

- Bounds for $|\mathcal{N}(S)_{jk}|$ from concentration inequality

BOUNDING $S(\hat{\Sigma} - \Sigma, l)$





- Not a regular blocking with size $2^{l-1} \log d$
- A regular blocking with size $2^{l-2} \log d$

Putting All Together

$$L \asymp \begin{cases} \log_2(d/\log d) & \text{if } d \le n^{1/(2\alpha+1)} \\ \log_2\left(n^{\frac{1}{2\alpha+1}}/\log d\right) & \text{if } \log d < n^{1/(2\alpha+1)} \text{ and } n^{1/(2\alpha+1)} \le d \\ \log_2(\log p/\log d) & \text{if } n^{1/(2\alpha+1)} \le \log d \end{cases}$$

$$\Rightarrow \quad \mathbb{E}\|\hat{\Sigma}-\Sigma\|^2 \lesssim \begin{cases} d/n & \text{if } d \le n^{1/(2\alpha+1)} \\ n^{-\frac{2\alpha}{2\alpha+1}} & \text{if } \log d < n^{1/(2\alpha+1)} \text{ and } n^{1/(2\alpha+1)} \le d \\ \log d/n & \text{if } n^{1/(2\alpha+1)} \le \log d \end{cases}$$

RECAP

MAIN TECHNICAL TOOLS

- Concentration inequality for blocks of sample covariance matrix

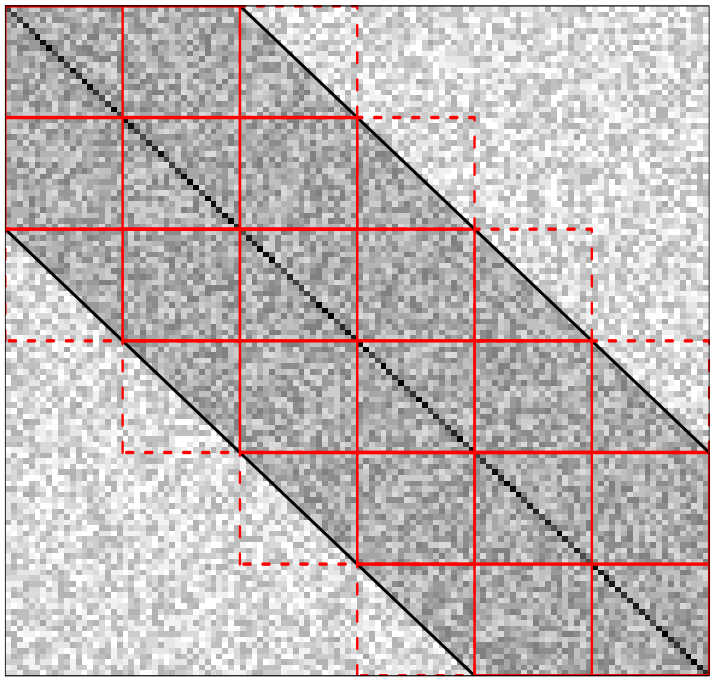
$$\|\bar{\Sigma}_B - \Sigma_B\| \lesssim_{\text{w.h.p.}} \sqrt{\frac{s(B) + \log d}{n}} \qquad \forall B \in \mathcal{B}$$

- Norm compression inequality

$$\text{Regular blocking} \Rightarrow \text{Norm compression transform} \Rightarrow \text{Norm expansion}$$

Could find use in other problems...

Optimality of Banding



- Blocking with size $k/2$ with $k \asymp n^{1/(2\alpha+1)}$
- Concentration inequality for blocks
- Concentration inequality for half blocks
- Norm compression inequality
- Banding is rate optimal – not adaptive

BEYOND NORMALITY

Elliptical Distribution

$$f(\boldsymbol{x};\mu,\Sigma)=|\Sigma|^{-\frac{1}{2}}g\left((\boldsymbol{x}-\mu)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\mu)\right)$$

- Normal distribution
- t distribution/Cauchy distribution
- Exponential Power/Laplace distribution
- Logistic distribution
- Symmetric Stable distribution
- .....

Question: How to estimate Σ?

CHALLENGES

- Sample covariance matrix?
  - $\bar{\sum}^{\text{Sample}}$ may not be well defined
  - Example: Cauchy distribution
- Heavy tails
  - Sub-Gaussianity Commonly assumed that

$$\mathbb{E}e^{cX^2} < \infty$$

  - Counterexample: t-distribution has polynomial tails

Solution: Rank correlation based methods.

KENDALL'S τ

- Observations – $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Pair comparison – (X_i, Y_i) vs (X_j, Y_j)
  - Concordant pair – $(X_i - X_j)(Y_i - Y_j) > 0$
  - Discordant pair – $(X_i - X_j)(Y_i - Y_j) < 0$
- The Kendall τ coefficient

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

- Key: For elliptical distribution

$$\tau = \frac{2}{\pi} \arcsin(\rho)$$

Rank-based Block Thresholding

- Estimate rank correlation matrix

$$\bar{T}^{\text{Sample}} = (\hat{\tau}_{ij})$$

- Convert to estimate of Σ

$$\bar{\Sigma} = \left(\sin(\frac{\pi}{2}\hat{\tau}_{ij})\right)$$

- Apply block-thresholding to get final estimate $\hat{\Sigma}$
- Adaptivity

$$\left.\begin{array}{r}\text{concentration inequality}\\ \text{norm compression inequality}\end{array}\right\} \Longrightarrow \sup_{\substack{P\in\mathcal{E}\\ \Sigma\in\mathcal{C}_\alpha}} \mathbb{E}\|\hat{\Sigma}-\Sigma\|^2 \lesssim \min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log d}{n}, \frac{d}{n}\right\}$$

SUMMARY

- Adaptive estimation for high dimensional covariance matrices is of great practical interests
- Adaptivity can be achieved with block thresholding
- Technical tools developed could be used for other purposes
- References
  - Adaptive Covariance Matrix Estimation Through Block Thresholding (with T.T. Cai)
  - Parameter Estimation for High Dimensional Meta-Elliptical Distributions (with Y. Xiao)
  - High Dimensional Semiparametric Gaussian Copula Graphical Models (with H. Liu et al.)