

High-dimensional Covariance Matrix Estimation

Jianqing Fan

Princeton University



With Yingying Fan, Jinchi Lv, Clifford Lam

<http://www.princeton.edu/~jqfan>

February 13, 2008

Outline

- Introduction
- Factor-based Estimating covariance matrix
 - ◆ Impact of dimensionality;
 - ◆ Applications to risk assessment and portfolio allocation.

Outline

- Introduction
- Factor-based Estimating covariance matrix
 - ◆ Impact of dimensionality;
 - ◆ Applications to risk assessment and portfolio allocation.
- Estimation of sparse covariance and precision matrices
 - ◆ sparsistency and rates of convergence;
 - ◆ graphical network

Introduction

Covariance matrix pervades all facets of sciences and humanities:

Portfolio Management: A portfolio of p assets with returns \mathbf{X} and allocation vector \mathbf{w} .

Introduction

Covariance matrix pervades all facets of sciences and humanities:

Portfolio Management: A portfolio of p assets with returns \mathbf{X} and allocation vector \mathbf{w} .

★ What is the risk of the portfolio: $\text{var}(\mathbf{w}^T \mathbf{X}) = \mathbf{w}^T \Sigma \mathbf{w}$?

★ What is the optimal allocation?

$$\min_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w}, \quad \text{s.t. } \mathbf{w}^T \mathbf{1} = 1, \text{ and } \mathbf{w}^T \boldsymbol{\mu} = .15?$$

$$\text{Solution: } \mathbf{w} = c_1 \Sigma^{-1} \boldsymbol{\mu} + c_2 \Sigma^{-1} \mathbf{1}$$

★ What is optimal sparse allocation?

Classification

- Disease classification using bioinformatic data.
- Document or text classification: E-mail spam.

Problem: ● n_1 data points from $N(\mu_1, \Sigma)$;

● n_2 points from $N(\mu_2, \Sigma)$.

What is the class label of X ?

Classification

■ Disease classification using bioinformatic data.

■ Document or text classification: E-mail spam

Problem: ● n_1 data points from $N(\mu_1, \Sigma)$;

● n_2 points from $N(\mu_2, \Sigma)$.

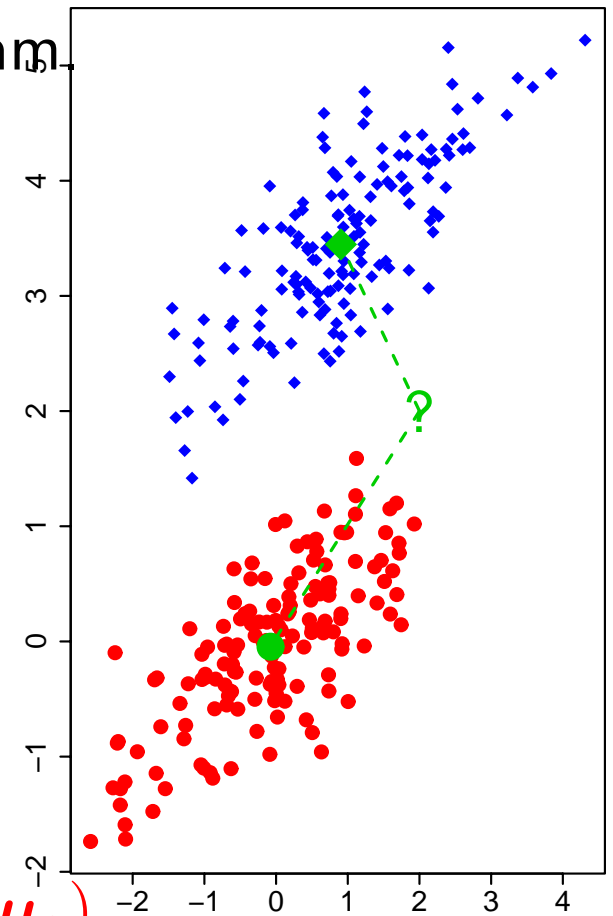
What is the class label of X ?

★ Fisher discr. $(\mathbf{x} - \mu)' \Sigma^{-1} (\mu_1 - \mu_2) > 0$

(nearest centroid in Mahalanobis distance):

★ optimal weighting on features: $\Sigma^{-1} (\mu_1 - \mu_2)$

★ Feature selection: what is sparse weighting?



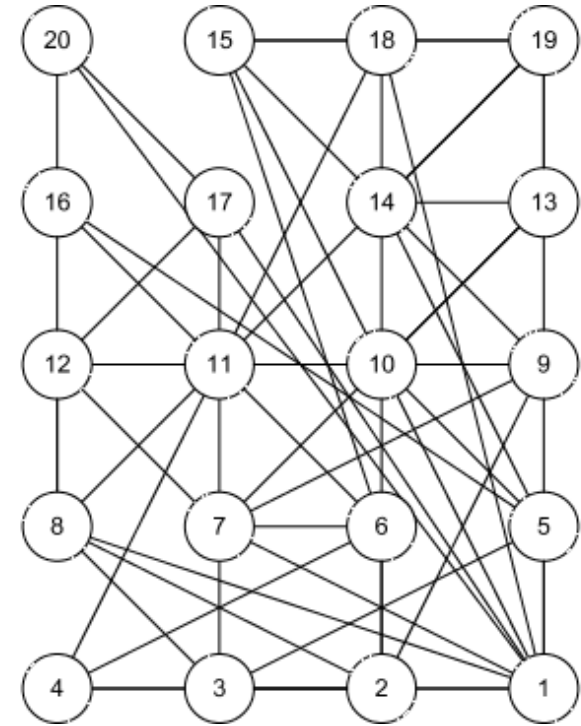
Graphical modeling

Graphic model:

- ★ Vertex: components of vector \mathbf{X}
- ★ Edges: the conditional dependence.

Precision matrix: $\Sigma^{-1} = (\sigma^{ij})$,

$$\sigma^{ij} = \text{cov}(X_i, X_j | \text{rest}).$$



- A simple network graph corresponds to a sparse precision matrix.

Classical Multivariate Analysis

$p : 3 \sim 8, n = 30 - 100$

Asymptotic framework: $n \rightarrow \infty$, but p fixed.

Classical Multivariate Analysis

$$p : 3 \sim 8, n = 30 - 100$$

Asymptotic framework: $n \rightarrow \infty$, but p **fixed**.

- ◆ inappropriate for many contemporary applications.
- ◆ more appropriate: $p \rightarrow \infty$ and $n \rightarrow \infty$ and examine the impact of dimensionality.

Challenge of High Dimensionality

- ♣ Estimating **high-dim** cov-matrices is intrinsically challenging.
 - Suppose we have 500 (**2000**) stocks to be managed.
 - There are 125K (**2 m**) free parameters!
 - Yet, 3-year daily returns yield only about sample size $n = 750$.

Challenge of High Dimensionality

- ♣ Estimating **high-dim** cov-matrices is intrinsically challenging.
- Suppose we have 500 (**2000**) stocks to be managed.
- There are 125K (**2 m**) free parameters!
- Yet, 3-year daily returns yield only about sample size $n = 750$.
- Accurately estimating it poses significant challenges.
- Impact of dim is large and poorly understood:

$$\text{Risk: } \mathbf{w}^T \hat{\Sigma} \mathbf{w}. \quad \text{Allocation: } \hat{c}_1 \hat{\Sigma}^{-1} \mathbf{1} + \hat{c}_2 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}.$$

Dimensionality reduction

Sample Covariance: ♦ Ill-conditioned; ♦ divergence of spectra

(Johnstone 01).

Dimensionality reduction: Impose some structure on covariance.

- Factor models: (Engle& Watson, 81; Chamberlain&Rothschild, 83; Diebold&Nerlove, 89; Aguilar&West, 00; Stock&Watson, 05; Fan, Fan and Lv, 07)

Dimensionality reduction

Sample Covariance: ♦ Ill-conditioned; ♦ divergence of spectra

(Johnstone 01).

Dimensionality reduction: Impose some structure on covariance.

- Factor models: (Engle & Watson, 81; Chamerlain & Rothschild, 83; Diebold & Nerlove, 89; Aguilar & West, 00; Stock & Watson, 05; Fan, Fan and Lv, 07)
- Sparsity Models. (Pourahmadi, 00, Boik, 03, Wu & Pourahmadi, 03; Huang, Liu, Pourahmadi, 04; Li and Gui 06, Bickel & Levina, 07; Rothman, *et al.*, 07, Wagaman and Levina, 07).

Seeking sparse solutions

Biologists: To understand molecular mechanisms and biochemical pathways.

Financial investors: To reduce transaction/monitoring costs and enhance portfolio performance.

Seeking sparse solutions

Biologists: To understand molecular mechanisms and biochemical pathways.

Financial investors: To reduce transaction/monitoring costs and enhance portfolio performance.

Social scientists: To understand simple and comprehensible social networks.

Seeking sparse solutions

Biologists: To understand molecular mechanisms and biochemical pathways.

Financial investors: To reduce transaction/monitoring costs and enhance portfolio performance.

Social scientists: To understand simple and comprehensible social networks.

Statisticians: To reduce noise accumulations and increase accuracy of parameters.

Seeking sparse solutions

Biologists: To understand molecular mechanisms and biochemical pathways.

Financial investors: To reduce transaction/monitoring costs and enhance portfolio performance.

Social scientists: To understand simple and comprehensible social networks.

Statisticians: To reduce noise accumulations and increase accuracy of parameters.

Numerical Analysis: To find stable and nearly optimal solutions.

Factor-model Based Estimation

Factor Model

- derived by Ross (76) using APT and Chamberlian and Rothchild (83) in large economies.
- an extension of CAPM

Factor model: K allows to depend on p

$$Y_i = b_{i1}f_1 + \cdots + b_{iK}f_K + \varepsilon_i, \quad i = 1, \cdots, p_n$$

Y_i — excess return of the i -th asset;

f_1, \cdots, f_K — factors that influence the returns.

Factor Model

- derived by Ross (76) using APT and Chamberlian and Rothchild (83) in large economies.
- an extension of CAPM

Factor model: K allows to depend on p

$$Y_i = b_{i1}f_1 + \cdots + b_{iK}f_K + \varepsilon_i, \quad i = 1, \cdots, p_n$$

Y_i —excess return of the i -th asset;

f_1, \cdots, f_K —factors that influence the returns.

◆ $\{\varepsilon_i\}$ are idiosyncratic noises, uncorrelated with the factors;

◆ reduce biases; dim-reduction: p/K .

Fama-French factor models

- 3-factor for stock portfolios
- 5-factor when bonds involved.
- CRSP value-weighted stock index; (all stocks in NYSE, AMEX, NASDAQ)
- difference of returns between large and small capitalization;
- difference of returns between high and low book-to-market ratios;

$$f_2 = 1/2(\text{Small Value} + \text{Big Value}) - 1/2(\text{Small Growth} + \text{Big Growth})$$

$$f_3 = 1/3(\text{SV} + \text{SN} + \text{SG}) - 1/3(\text{BV} + \text{BN} + \text{BG}).$$

Fama-French factor models

- 3-factor for stock portfolios
- 5-factor when bonds involved.
- CRSP value-weighted stock index; (all stocks in NYSE, AMEX, NASDAQ)
- difference of returns between large and small capitalization;
- difference of returns between high and low book-to-market ratios;
$$f_2 = 1/2(\text{Small Value} + \text{Big Value}) - 1/2(\text{Small Growth} + \text{Big Growth})$$
$$f_3 = 1/3(\text{SV} + \text{SN} + \text{SG}) - 1/3(\text{BV} + \text{BN} + \text{BG}).$$
- ◆ a term structure factor (yield spread between long and short bonds);
- ◆ a default risk (yield spread high and low grade bonds)

International Diversification

Emerging Market: $n = 2,000$ investible.

Model: $r = g + C + I + S + \varepsilon$

- g = the return of EM universe
 - C and I are country and industry
 - S are style factors (size, BP).
-
- How to allocate the portfolio efficiently, robustly, and sparsely?
 - How to choose factors?

Model-based estimation

Matrix form — multiperiod $Y_{ti} = b_{i1}f_{t,1} + \cdots + b_{iK}f_{t,K} + \varepsilon_{t,i}$

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \varepsilon_t. \quad \text{Factors } \mathbf{f}_t \text{ observable}$$

Covariance structure: $\Sigma = \mathbf{B}\text{var}(\mathbf{f})\mathbf{B}^T + \Sigma_0$

Remarks: Σ_0 diagonal, K can depend on p .

Model-based estimation

Matrix form — multiperiod $Y_{ti} = b_{i1}f_{t,1} + \cdots + b_{iK}f_{t,K} + \varepsilon_{t,i}$

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \varepsilon_t. \quad \text{Factors } \mathbf{f}_t \text{ observable}$$

Covariance structure: $\Sigma = \mathbf{B}\text{var}(\mathbf{f})\mathbf{B}^T + \Sigma_0$

Remarks: Σ_0 **diagonal**, K can depend on p .

Estimated covariance: $\hat{\Sigma}$ — regression + substitution.

Sample covariance matrix: $\hat{\Sigma}_{\text{sam}}$

Questions

- How do the estimation errors **grow with** p_n and K_n ?
- How large can p_n be such that the error in estimated covariance is **negligible** in portfolio allocation and risk assessment?

Questions

- How do the estimation errors **grow with** p_n and K_n ?
- How large can p_n be such that the error in estimated covariance is **negligible** in portfolio allocation and risk assessment?
- How does the model based estimator **perform** compared with the sample one?
- Under which situations can the model-based approach gain **substantially / marginally**?

Convergence rates

Theorem 1: $\max \left| \lambda_k(\widehat{\Sigma}) - \lambda_k(\Sigma) \right| = o_P\{\mathbf{K}_n \mathbf{p}_n (\log n/n)^{1/2}\}$.

—the **same rate** as that from the sample covariance.

■the rate **optimality** can be seen a toy example:

$$\mathbf{B} = \mathbf{1} \text{ and } \text{var}(\varepsilon) = I_{p_n} \implies \Sigma = \text{var}(f) \mathbf{1}\mathbf{1}^T + I_{p_n}.$$

— $\|\widehat{\Sigma} - \Sigma\|^2 = p_n^2 |\widehat{\text{var}}(f) - \text{var}(f)|^2$.

—Engen-value: $(1 + \text{var}(f)p_n), 1, \dots, 1$.

Conclusion: ♠ does not help on estimating Σ and hence risks.

Choice of Norms

Frobenius norm: $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^T \mathbf{A}) = \sum_{i,j} a_{i,j}^2$

Operator norm: $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}^T \mathbf{A})$.

Quadratic norm: $\|\hat{\Sigma} - \Sigma\|_Q^2 = \text{tr}[\hat{\Sigma}\Sigma^{-1} - I]^2/p$.

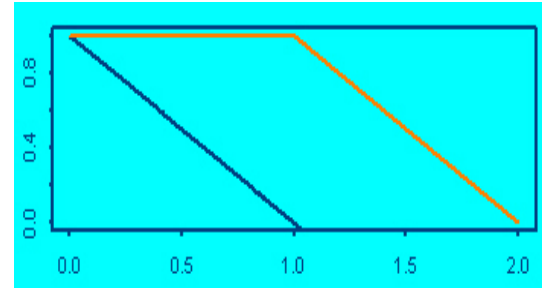
Entropy loss: $\text{tr}(\hat{\Sigma}\Sigma^{-1}) - \log |\hat{\Sigma}\Sigma^{-1}| - p$

Strength of factor structure

Theorem 2: If $p_n = n^\alpha$, $K_n = O(p_n^{\alpha_1/\alpha})$, then

$$\|\hat{\Sigma} - \Sigma\|_Q = O_P(n^{-\beta/2})$$

$$\|\hat{\Sigma}_{\text{sam}} - \Sigma\|_Q = O_P(n^{-\beta_1/2})$$

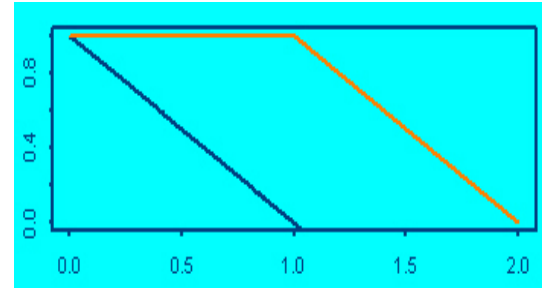


Strength of factor structure

Theorem 2: If $p_n = n^\alpha$, $K_n = O(p_n^{\alpha_1/\alpha})$, then

$$\|\hat{\Sigma} - \Sigma\|_Q = O_P(n^{-\beta/2})$$

$$\|\hat{\Sigma}_{\text{sam}} - \Sigma\|_Q = O_P(n^{-\beta_1/2})$$



$$\beta = \min(1 - 2\alpha_1, 2 - \alpha - \alpha_1), \quad \beta_1 = 1 - \max(\alpha, 3\alpha_1/2, 3\alpha_1 - \alpha).$$

Theorem 3: Under the Frobenius norm,

$$\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|^2 = O_P(\mathbf{p}_n^2 \mathbf{K}_n^4 \log n/n)$$

◆ an order $(p_n/K_n)^2$ **smaller** than $\|\hat{\Sigma}_{\text{sam}}^{-1} - \Sigma^{-1}\|^2$.

Summary: For estimating Σ^{-1} , the factor model gains.

Impact on portfolio allocation

Mean-variance optimality: Markowitz (1952)

$$\min_{\xi' \mathbf{1}=1, \xi' \boldsymbol{\mu}=a} \xi' \boldsymbol{\Sigma} \xi, \quad \text{Sol : } \xi = c_1 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + c_2 \boldsymbol{\Sigma}^{-1} \mathbf{1}.$$

Theorem 5 (optimal portfolio) Using the factor model,

$$\left| \widehat{\boldsymbol{\xi}}_n' \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}'_n \boldsymbol{\Sigma}_n \boldsymbol{\xi}_n \right| = o((\mathbf{p}_n \mathbf{K}_n)^2 (\log n/n)^{1/2}),$$

whereas using the sample covariance, the rate is (p_n/K_n) **worse**.

■ The same results apply to global minimum portfolio.

Impact on Risk Assessment

Volatility: of a portfolio with weights \mathbf{w}_0 is $\mathbf{w}_0^T \widehat{\Sigma} \mathbf{w}_0$, associated with risk management.

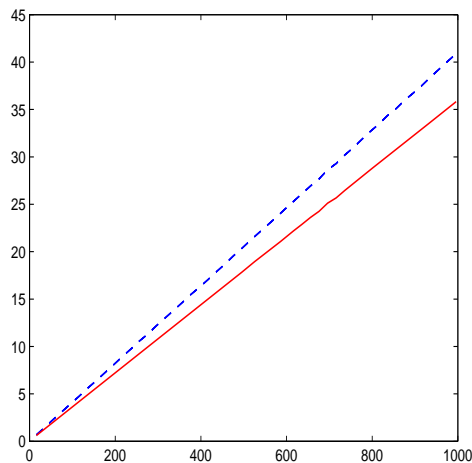
Theorem 6: Given a portfolio $\mathbf{1}^T \mathbf{w}_0 = 1$, we have

$$\left| \mathbf{w}_0^T \widehat{\Sigma} \mathbf{w}_0 - \mathbf{w}_0^T \Sigma \mathbf{w}_0 \right| = o_P \left\{ (\mathbf{p}_n^2 \mathbf{K}_n) (\log n / n)^{1/2} \right\}.$$

- the **same rate** as that from the sample covariance;
- If no short position, rate is $o_P \left\{ \mathbf{p}_n \mathbf{K}_n (\log n / n)^{1/2} \right\}$.

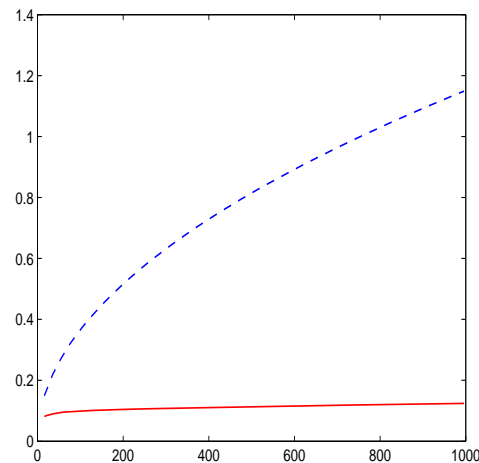
Comparison of Performance

- Use Fama-French 3-factor models; calibrated to market data.
- Examine impact of dimensionality based 500 simulations
- factor-based covariance matrix and sample one.

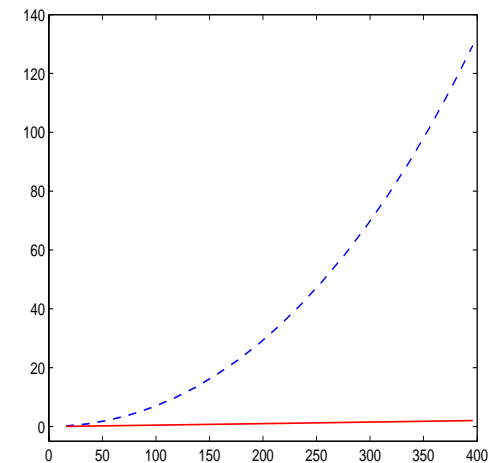


Frobenius norm

$$\text{tr}(\hat{\Sigma} - \Sigma)^2$$



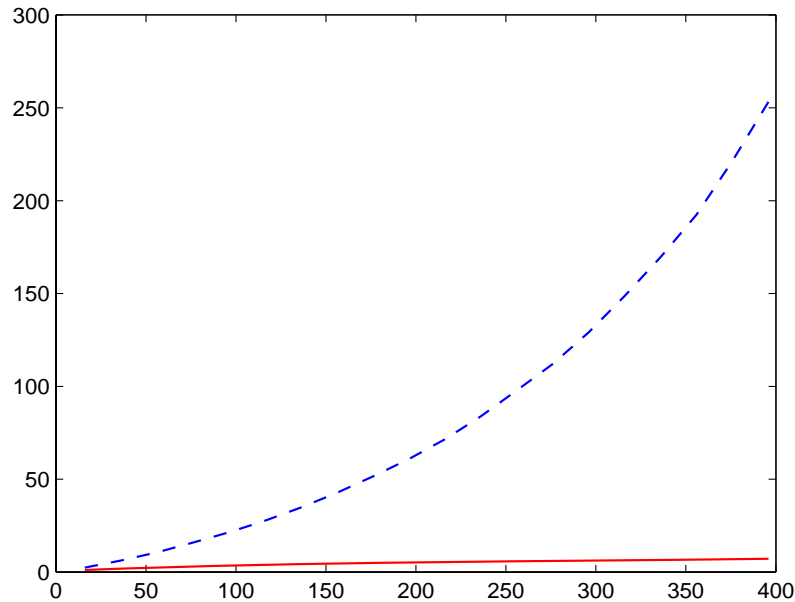
$$\|\cdot\|_{\Sigma} \\ \text{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2$$



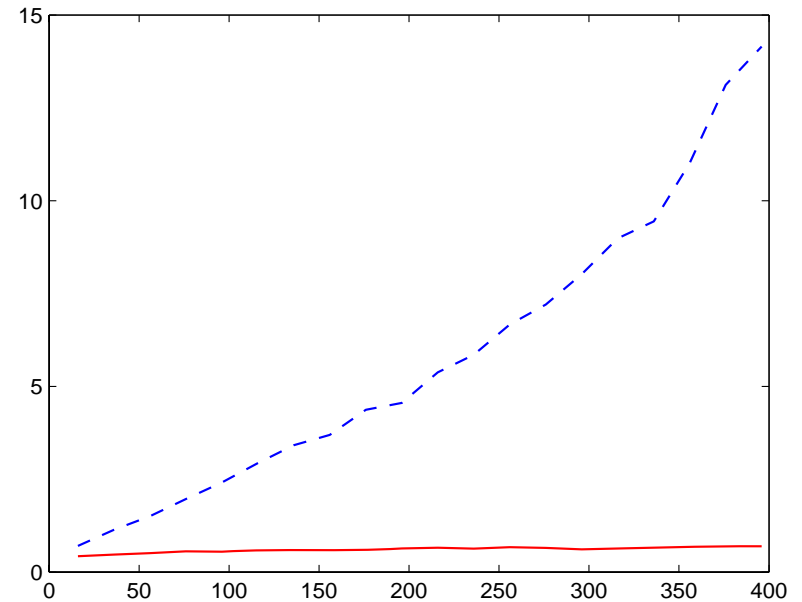
entropy loss

$$\text{tr}(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p$$

Estimation of Σ^{-1} under Frobenius norm



$$E\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F^2$$



SD

- Also examined the impact on portfolio choice and risk assessment
- consistent with theory.

What shall we do?

Equi-loadings: ■ approximately the same within a homogeneous group (e.g. sectors, country). ■ vary smoothly over time

— Loadings can be estimated more precisely and hence rates of convergence improved.

What shall we do?

Equi-loadings: ■ approximately the same within a homogeneous group (e.g. sectors, country). ■ vary smoothly over time
— Loadings can be estimated more precisely and hence rates of convergence improved.

Sparse utility maximization: With $|w|$ being L_0 or L_1 -norm.

$$\max_{|w| \leq \delta} (1 - \mathbf{w}^T \mathbf{1}) r_0 + \mathbf{w}^T \boldsymbol{\mu} - \frac{\lambda}{2} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}.$$

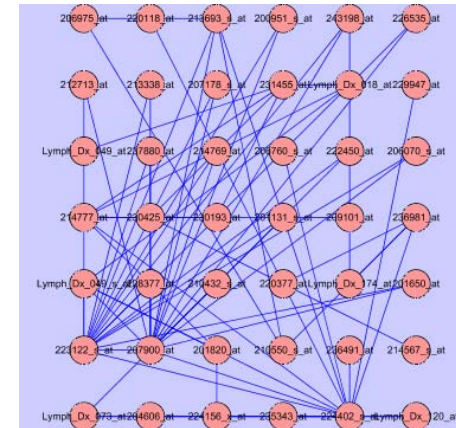
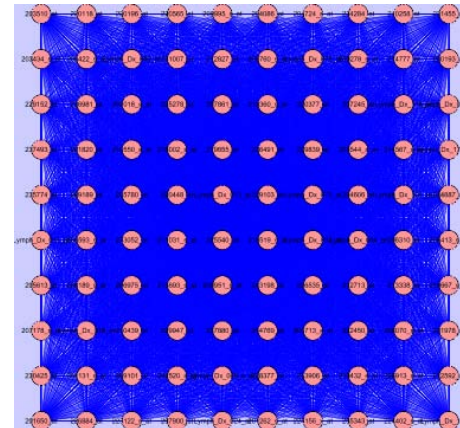
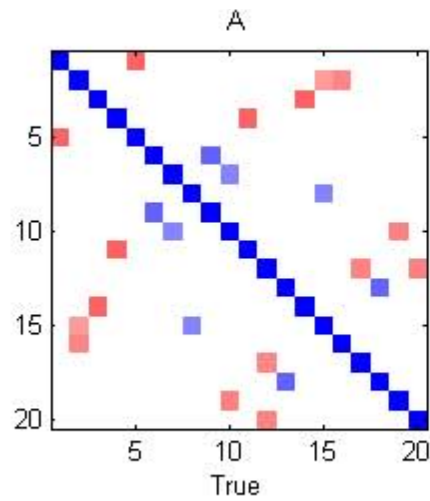
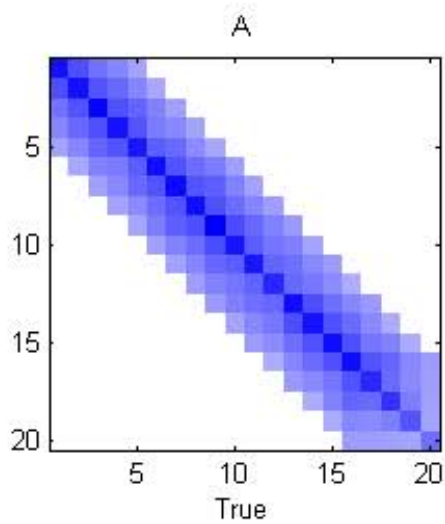
- ★ closely related to regression problem with L_1 -penalty.
- ★ LARS can be used along with newly estimated covariance.

Sparsity-Based Estimation

Sparsity

■ Covariance matrix Σ (Bickel & Levina, 07; El Karoui, 07; Yuan & Li)

■ Precision matrix Σ^{-1} (Meinshausen, Bühlman, 06; Rothman, et al, 07; Wagaman Levina, 07))



■ Sparse Modified Cholesky Decomposition (Pourahmadi, 99; Huang, et al, 06)

Penalized likelihood

■ Sparsity in covariance (Fan & Li, 01; Rothman, et al, 07; Lam & Fan, 07)

$$q_1(\boldsymbol{\Sigma}) = \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \log |\boldsymbol{\Sigma}| + \sum_{i \neq j} p_{\lambda_{n1}}(|\sigma_{ij}|),$$

— \mathbf{S} = sample covariance matrix

■ Sparsity in precision matrix (Fan & Li, 01; Rothman, et al, 07; Lam & Fan, 07)

$$q_2(\boldsymbol{\Omega}) = \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \log |\boldsymbol{\Omega}| + \sum_{i \neq j} p_{\lambda_{n2}}(|\omega_{ij}|),$$

Penalized likelihood

■ Sparsity in covariance (Fan & Li, 01; Rothman, et al, 07; Lam & Fan, 07)

$$q_1(\mathbf{\Sigma}) = \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) + \log |\mathbf{\Sigma}| + \sum_{i \neq j} p_{\lambda_{n1}}(|\sigma_{ij}|),$$

— \mathbf{S} = sample covariance matrix

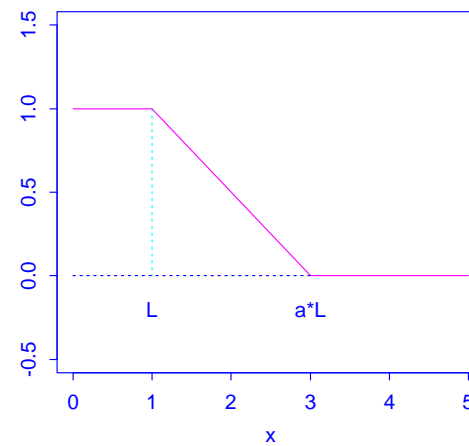
■ Sparsity in precision matrix (Fan & Li, 01; Rothman, et al, 07; Lam & Fan, 07)

$$q_2(\mathbf{\Omega}) = \text{tr}(\mathbf{S}\mathbf{\Omega}) - \log |\mathbf{\Omega}| + \sum_{i \neq j} p_{\lambda_{n2}}(|\omega_{ij}|),$$

Choice of penalty function:

★ L_1 : $p_{\lambda}(|x|) = \lambda|x|$;

★ SCAD



Computation

Penalized L_1 : Convex optimization s.t. positivity constraint:

■ d'Aspremont et al (07) two first-order semi-definite programming algorithms

■ MAXDET algorithm (Vandenberghe et al, 1998; Yuan and Lin 07)

Computation

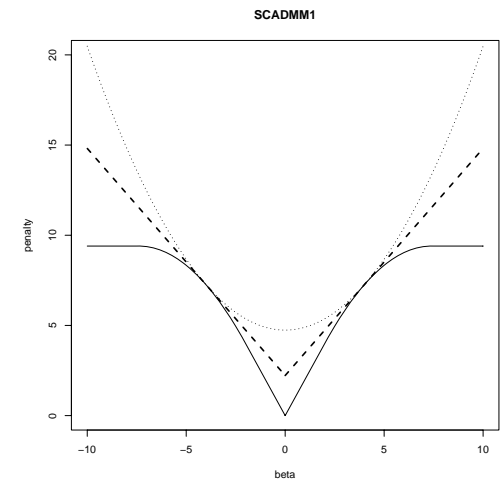
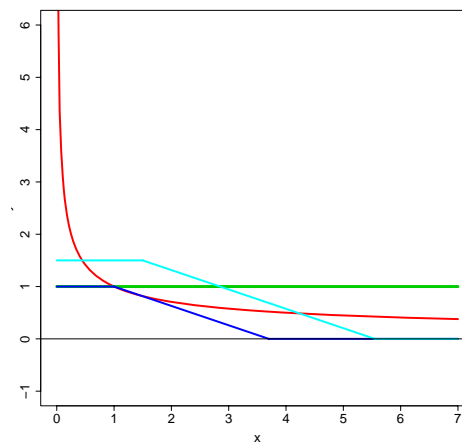
Penalized L_1 : Convex optimization s.t. positivity constraint:

■ d'Aspremont et al (07) two first-order semi-definite programming algorithms

■ MAXDET algorithm (Vandenberghe et al, 1998; Yuan and Lin 07)

SCAD penalty: LLA approximation (Zou & Li, 08; Fan, et al.08):

$$\min_{\Omega > 0} \text{tr}(\mathbf{S}\Omega) - \log |\Omega| + \sum_{i \neq j} p'_{\lambda n_2}(|\omega_{ij}^{(k)}|)(|\omega_{ij}|).$$



Computation

Penalized L_1 : Convex optimization s.t. positivity constraint:

■ d'Aspremont et al (07) two first-order semi-definite programming algorithms

■ MAXDET algorithm (Vandenberghe et al, 1998; Yuan and Lin 07)

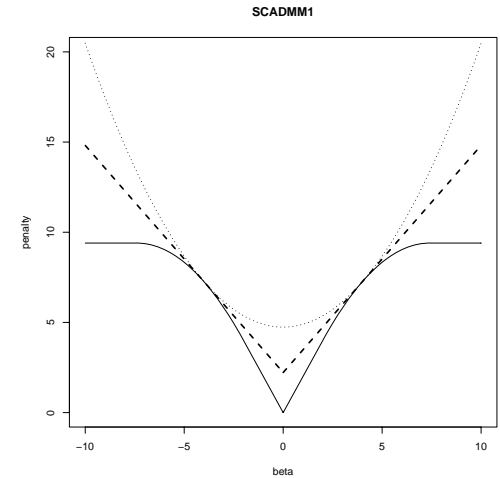
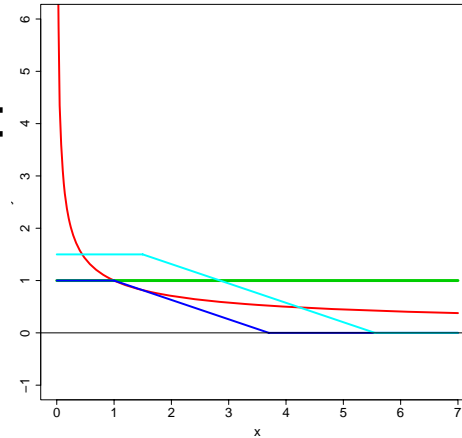
SCAD penalty: LLA approximation (Zou & Li, 08; Fan, et al.08):

$$\min_{\Omega > 0} \text{tr}(\mathbf{S}\Omega) - \log |\Omega| + \sum_{i \neq j} p'_{\lambda n_2}(|\omega_{ij}^{(k)}|)(|\omega_{ij}|).$$

Algorithmic Convergence:

$$q_2(\Omega^{(k+1)}) \geq q_2(\Omega^{(k)})$$

(Fan, Feng, Wu, 07)



Rates of Convergence for Σ

Let $S_1 = \{(i, j) : \sigma_{ij,0} \neq 0\}$ and $s_{n1} = |S_1| - p$,

Theorem 1: If Ω_0 is well conditioned, $(p_n + s_{n1}) \log p_n/n = o(1)$

and $\lambda_{n1}^2 \succeq (s_{n1} + 1) \log p_n/n$,

$$\|\hat{\Sigma} - \Sigma_0\|_{\mathbf{F}}^2 = \mathbf{O}_{\mathbf{P}}\{(p_n + s_{n1}) \log p_n/n\},$$

provided that $\min_{(i,j) \in S_1} |\sigma_{ij}^0|/\lambda_{n1} \rightarrow \infty$

Rates of Convergence for Σ

Let $S_1 = \{(i, j) : \sigma_{ij,0} \neq 0\}$ and $s_{n1} = |S_1| - p$,

Theorem 1: If Ω_0 is well conditioned, $(p_n + s_{n1}) \log p_n/n = o(1)$

and $\lambda_{n1}^2 \succeq (s_{n1} + 1) \log p_n/n$,

$$\|\hat{\Sigma} - \Sigma_0\|_F^2 = \mathbf{O}_P\{(p_n + s_{n1}) \log p_n/n\},$$

provided that $\min_{(i,j) \in S_1} |\sigma_{ij}^0|/\lambda_{n1} \rightarrow \infty$

$$a_{n1} = \max_{(i,j) \in S_1} p'_{\lambda_{n1}}(|\sigma_{ij}^0|) = \begin{cases} \lambda_{n1}, & L_1\text{-penalty} \\ 0, & \text{SCAD-penalty} \end{cases}$$

$$= O(\{1 + p_n/(s_{n1} + 1)\}(\log p_n/n)^{1/2}).$$

■ For L_1 , local minimizer becomes global one. Bias needs controlling.

Sparsistency for Σ

Sparsistency: all parameters that are zero are actually estimated as zero with probability tending to one (Ravikumar, et al, 07).

Theorem 2: If $\lambda_{n1}^2 \succeq (p_n + s_{n1}) \log p_n/n$, then

$\hat{\sigma}_{ij} = 0$ for all $(i, j) \in S_1^c$, with probability tending to 1.

Sparsistency for Σ

Sparsistency: all parameters that are zero are actually estimated as zero with probability tending to one (Ravikumar, et al, 07).

Theorem 2: If $\lambda_{n1}^2 \succeq (p_n + s_{n1}) \log p_n/n$, then

$\hat{\sigma}_{ij} = 0$ for all $(i, j) \in S_1^c$, with probability tending to 1.

♠ For L_1 -penalty, conditions compatible only when $s_{n1} = O(p_n^{1/2})$.

♠ The bias induced by L_1 undermines the reliable choice of λ_{n1} .

♠ Derived the asymptotic normality for non-zero elements.

Estimating Correlation Matrix

- Sparsity structure is consistent with that of covariance

- Correlation matrix can be estimated by minimizing

$$\text{tr}(\mathbf{\Gamma}^{-1}\hat{\mathbf{\Gamma}}) + \log |\mathbf{\Gamma}| + \sum_{i \neq j} p_{\nu_{n1}}(|\gamma_{ij}|).$$

— $\mathbf{\Gamma}$ = sample correlation

- An alternative estimator $\tilde{\Sigma}$ can be constructed.

Results for Estimating Correlation matrix

Theorem 3: If $p_n/n = o(1)$, $s_{n1} \log p_n/n = o(1)$ and $\nu_{n1}^2 \asymp (s_{n1} + 1) \log p_n/n$, then

$$\|\hat{\Gamma} - \Gamma_0\|_{\mathbf{F}}^2 = \mathbf{O}_{\mathbf{P}}(s_{n1} \log p_n/n)$$

Results for Estimating Correlation matrix

Theorem 3: If $p_n/n = o(1)$, $s_{n1} \log p_n/n = o(1)$ and $\nu_{n1}^2 \asymp (s_{n1} + 1) \log p_n/n$, then

$$\|\hat{\Gamma} - \Gamma_0\|_{\mathbf{F}}^2 = \mathbf{O}_{\mathbf{P}}(s_{n1} \log p_n/n)$$

$$\|\tilde{\Sigma} - \Sigma_0\|^2 = O_P\{(s_{n1} + 1) \log p_n/n\}.$$

■ Sparsistency also for correlation matrix.

■ The conditions are compatible only when $s_{n1} = O(1)$.

Estimation Sparse Precision Matrix Ω

■ Parameterized by $\Omega = \Sigma^{-1}$ instead of Σ

Rates of Convergence: If $(p_n + s_{n2}) \log p_n/n = o(1)$, and bias is negligible, and $\lambda_{n2}^2 \succeq (s_{n2} + 1) \log p_n/n$, then

$$\|\hat{\Omega} - \Omega_0\|_F^2 = O_P\{(p_n + s_{n2}) \log p_n/n\}$$

Estimation Sparse Precision Matrix Ω

■ Parameterized by $\Omega = \Sigma^{-1}$ instead of Σ

Rates of Convergence: If $(p_n + s_{n2}) \log p_n/n = o(1)$, and bias is negligible, and $\lambda_{n2}^2 \succeq (s_{n2} + 1) \log p_n/n$, then

$$\|\hat{\Omega} - \Omega_0\|_F^2 = O_P\{(p_n + s_{n2}) \log p_n/n\}$$

Sparsistency: If $\lambda_{n2}^2 \succeq (p_n + s_{n2}) \log p_n/n$, then

$$\hat{\omega}_{ij} = 0 \text{ for all } (i, j) \in S_2^c.$$

Asymptotic normality: derived nonsparse elements.

Estimation Sparse Inverse Precision Matrix Ω

- Normalize: $\text{tr}(\Psi \hat{\Gamma}) - \log |\Psi| + \sum_{i \neq j} p_{\nu_{n2}}(|\psi_{ij}|)$
- Transform: $\tilde{\Omega} = \hat{W}^{-1} \hat{\Psi} \hat{W}^{-1}$

Rates of Convergence: if $(s_{n2} + 1) \log p_n/n = o(1)$ and

$\nu_{n2}^2 \succeq (s_{n2} + 1) \log p_n/n$, then

$$\|\hat{\Psi} - \Psi_0\|_F^2 = O_P((s_{n2} + 1) \log p_n/n)$$

Estimation Sparse Inverse Precision Matrix Ω

- Normalize: $\text{tr}(\Psi \hat{\Gamma}) - \log |\Psi| + \sum_{i \neq j} p_{\nu_{n2}}(|\psi_{ij}|)$
- Transform: $\tilde{\Omega} = \hat{W}^{-1} \hat{\Psi} \hat{W}^{-1}$

Rates of Convergence: if $(s_{n2} + 1) \log p_n/n = o(1)$ and

$\nu_{n2}^2 \succeq (s_{n2} + 1) \log p_n/n$, then

$$\|\hat{\Psi} - \Psi_0\|_F^2 = O_P((s_{n2} + 1) \log p_n/n)$$

$$\|\tilde{\Omega} - \Omega_0\|^2 = O_P((s_{n2} + 1) \log p_n/n).$$

Sparsistency: holds under the same condition.

- For L_1 penalty, the conditions are compatible when $s_{n2} = O(1)$.

Puzzle on improved rates of convergence

■ Ω^{-1} does not have known diagonals.

■ Explanations: ★ if $s_{n2} \succeq p_n$, it is trivial.

★ If $s_{n2} = o(p_n)$, most of off-diagonals are zero. At most $O(s_{n2})$

non-trivial columns:

$$\begin{pmatrix} 1 & & & \dots \\ & 1 & & \dots \\ & & 1 & \dots \\ & & \times & 1 & \dots \\ & & \vdots & \vdots & \ddots \end{pmatrix}$$

Ω^{-1} contains at most $O(s_{n2})$ unknown diagonals.

Extension to sparse Modified Cholesky

Innovation Algorithm: $y_i = \sum_{j=1}^{i-1} \phi_{ij} y_j + \epsilon_i$, which implies

$\mathbf{T}\Sigma\mathbf{T}^T = \mathbf{D}$, \mathbf{T} lower triangular with unit diagonal elements.

- Criterion: Least-squares $\text{tr}(\mathbf{T}^T \mathbf{T} \mathbf{S})$ or maximum likelihood or normalized maximum likelihood $\text{tr}(\mathbf{T}^T \mathbf{T} \hat{\mathbf{\Gamma}}) - 2 \log |\mathbf{T}|$.

Extension to sparse Modified Cholesky

Innovation Algorithm: $y_i = \sum_{j=1}^{i-1} \phi_{ij} y_j + \epsilon_i$, which implies

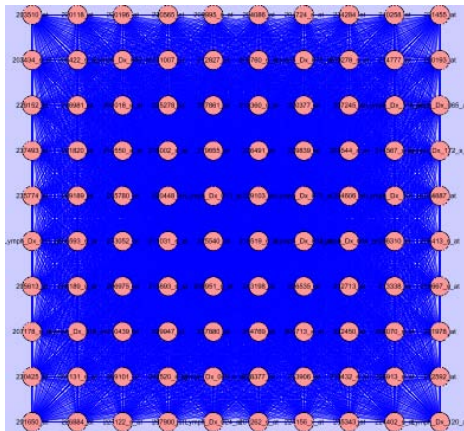
$\mathbf{T}\Sigma\mathbf{T}^T = \mathbf{D}$, \mathbf{T} lower triangular with unit diagonal elements.

- Criterion: Least-squares $\text{tr}(\mathbf{T}^T \mathbf{T} \mathbf{S})$ or maximum likelihood or normalized maximum likelihood $\text{tr}(\mathbf{T}^T \mathbf{T} \hat{\mathbf{\Gamma}}) - 2 \log |\mathbf{T}|$.
- ♠ Results: Rates of convergence, sparsistency.
- ♠ Normalized version improved rates of convergence.
- ♠ For L_1 , optimal rates and sparsistency impose the constraint $s_{n3} = O(p_n^{1/2})$ for ML and $s_{n3} = O(1)$ for normalized one.

Numerical Studies

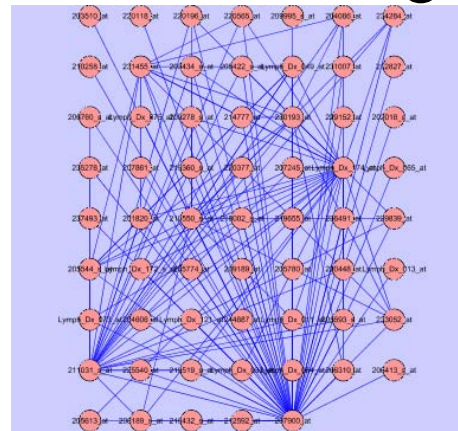
Burkitt Lymphoma Data

- Gene expressions from $n = 233$ diffuse large-B-cell lymphoma patients
- $p = 100$ genes with largest variance (Bernejee et al, 07)
- 6-fold cross-validation to select tuning parameters



LASSO

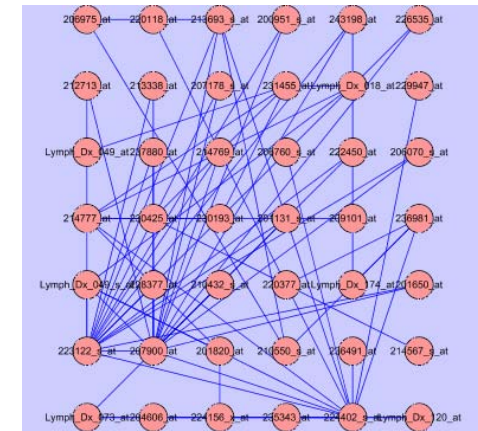
4947



Adaptive LASSO

147

Isolated nodes are omitted.



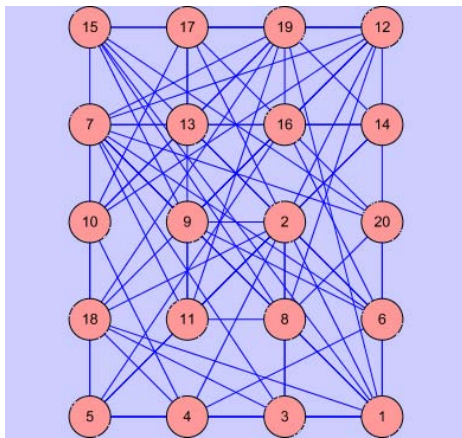
SCAD

68 edges

- Are edges in LASSO spurious? Lack of sparsistency?

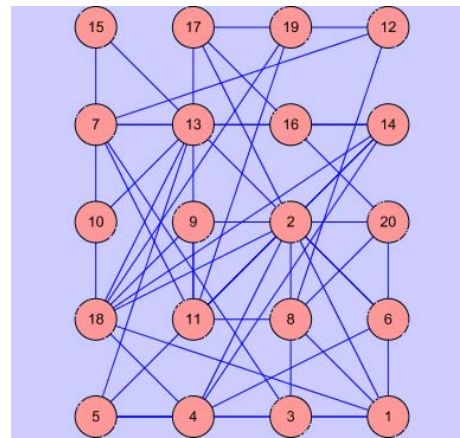
Lung Cancer Data

- Gene expressions from $n = 139$ lung adenocarcinomas
- $p = 80$ genes with largest t -statistic (Bhattacharjee et al, 01)



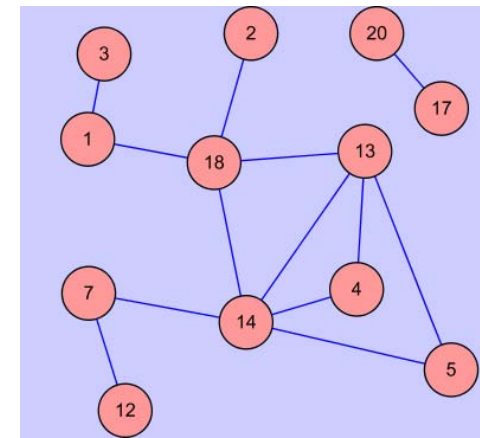
LASSO

1143



Adaptive LASSO

584



SCAD

126 edges

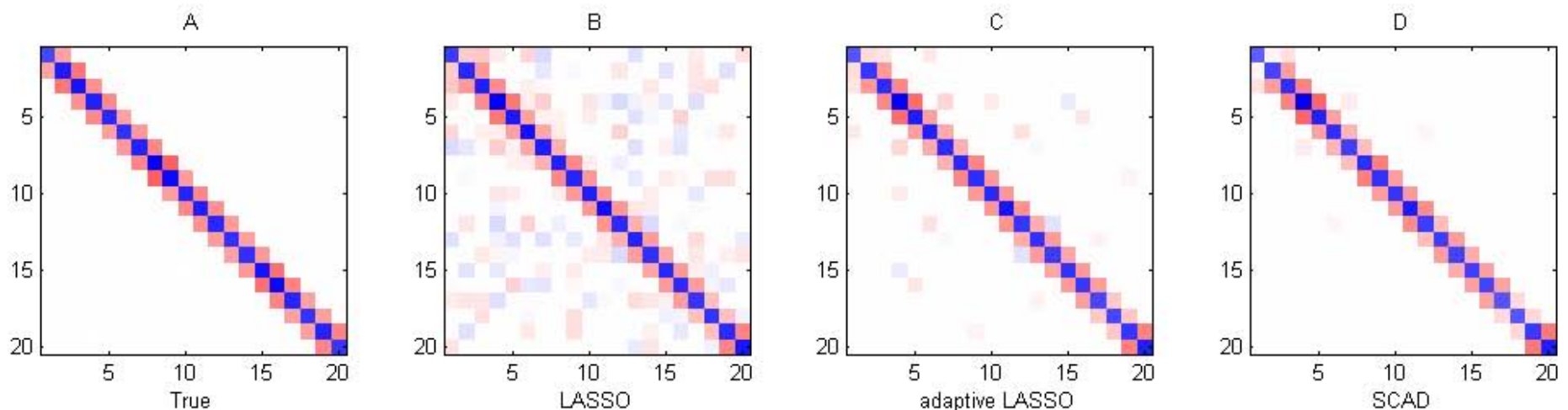
Isolated nodes omitted; only first 20 gene's network presented

- Are edges in LASSO spurious?

Simulated Example 1—Tri-diagonal

■ $n = 120, p = 20$; No. of Sim. = 100

■ $\Sigma = \exp(-a|s_i - s_j|)$ with $s_i - s_i \sim_{i.i.d} \text{Unif}(.5, 1)$.

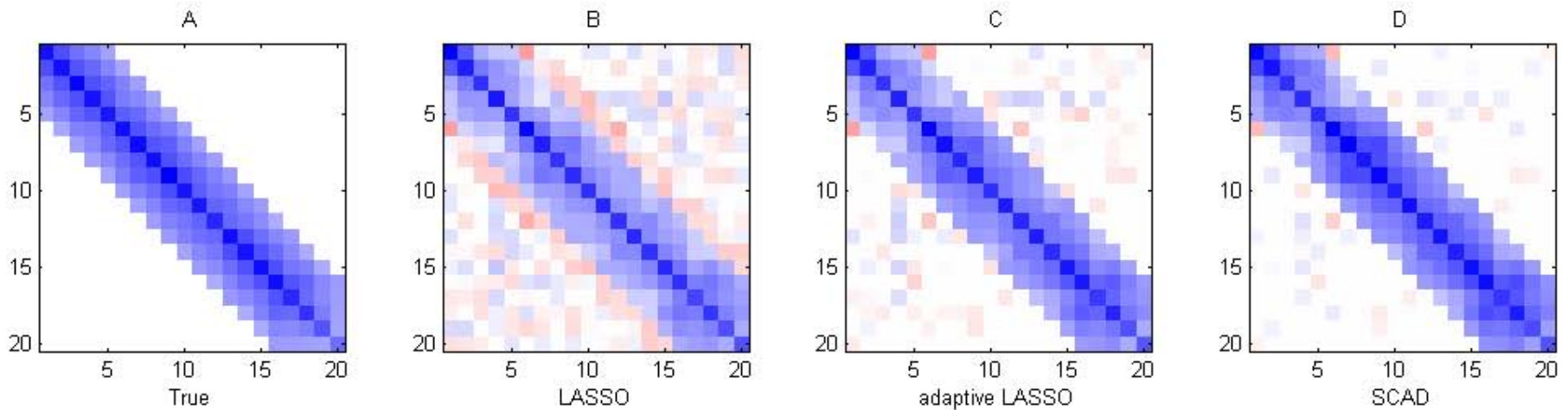


| | zero ₁ | zero ₂ | loss ₁ | loss ₂ |
|---------|-------------------|---------------------|-------------------|-------------------|
| LASSO | 0.0(0.0) | 127.2 (17.7) | 0.829(0.132) | 2.036(1.771) |
| A LASSO | 0.4(1.0) | 44.90(9.6) | 0.621(0.147) | 0.550(0.707) |
| SCAD | 0.4(1.1) | 41.3(13.4) | 0.699(0.179) | 1.330(1.568) |

Simulated Example 2 — Band matrix

■ $n = 120, p = 20$; No. of Sim. = 100

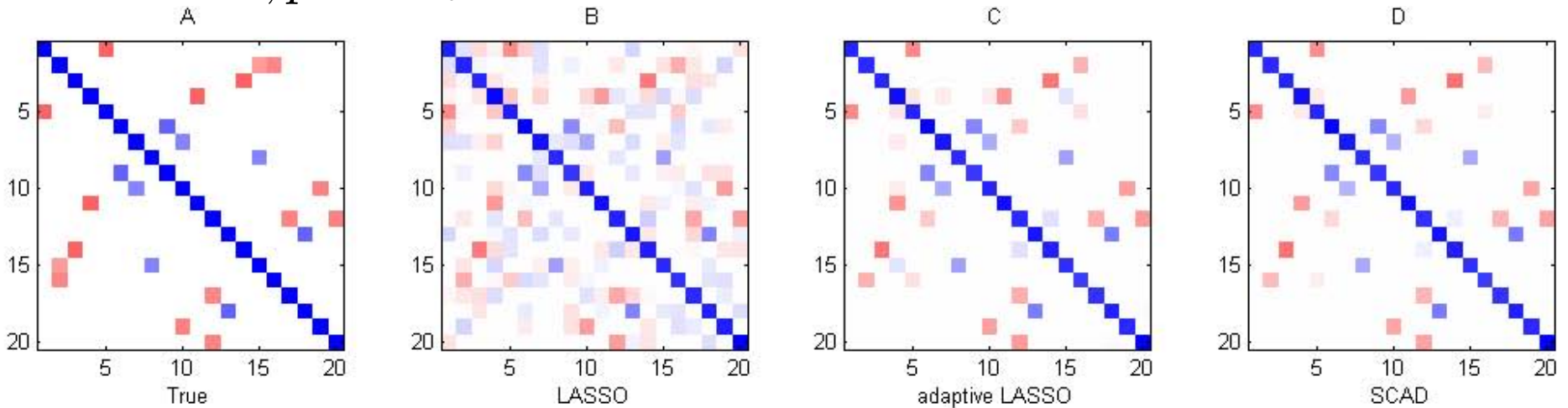
■ $\Omega = \mathbf{B}\mathbf{B}^T$ with $B_{ij} \sim_{i.i.d} \text{Unif}(.3, .5), |i - j| \leq 4$.



| | zero ₁ | zero ₂ | loss ₁ | loss ₂ |
|----------------|-------------------|--------------------|-------------------|-------------------|
| LASSO | 0.8(1.2) | 189.9 (9.4) | 1.855(0.196) | 2.007(2.072) |
| adaptive LASSO | 0.9(1.6) | 84.2(13.7) | 1.376(0.196) | 0.684(1.010) |
| SCAD | 1.0(1.6) | 54.7(13.1) | 1.244(0.229) | 3.219(2.352) |

Simulated Example 3 — General case

■ $n = 120, p = 20$; No. of Sim. = 100



| | zero ₁ | zero ₂ | loss ₁ | loss ₂ |
|---------|-------------------|---------------------|-------------------|-------------------|
| LASSO | 0.0(0.4) | 105.6 (19.8) | 0.6906(0.096) | 2.851(2.069) |
| A LASSO | 0.620(1.2) | 37.5(9.2) | 0.499(0.1179) | 0.569(0.752) |
| SCAD | 0.6(1.2) | 47.0(17.1) | 0.552(0.137) | 0.686(0.909) |

Conclusions — Model Based Estimation

- We propose and study the use of the factor model for estimating high-dim covariance matrix;
- When the dimensionality is high, the factor based estimator
 - ◆ significantly **outperforms** the sample covariance particularly for its inverse.

Conclusions — Model Based Estimation

- We propose and study the use of the factor model for estimating high-dim covariance matrix;
- When the dimensionality is high, the factor based estimator
 - ◆ significantly **outperforms** the sample covariance particularly for its inverse.
 - ◆ significantly **outperforms** the sample covariance in portfolio allocation;
 - ◆ **doesn't** improve the performance of risk assessment.

Conclusions — Sparsity Based Estimation

- Established rates of convergence, sparsistency and asymptotic normality: ★ Sparsistency requires a lower bound on λ .
- The cost of dimensionality is merely $\log p_n$.

Conclusions — Sparsity Based Estimation

- Established rates of convergence, sparsistency and asymptotic normality: ★ Sparsistency requires a lower bound on λ .
- The cost of dimensionality is merely $\log p_n$.
- Better rates of convergence is obtained when estimate sparse correlation or inverse correlation matrix.
- Penalized L_1 induces biases and is hard to be optimal and sparsistency.

| | Optimal rates | Sparsistency | L_1 -sparsistency |
|-----------------|--|--|----------------------|
| Cov / Precision | $\left(\frac{s_n \log p_n}{n}\right)^{1/2}$ | $\lambda_n \succ \left(\frac{s_n \log p_n}{n}\right)^{1/2}$ | $s_n = O(p_n^{1/2})$ |
| Corr/Inv. Corr | $\left(\frac{s'_n \log p_n}{n}\right)^{1/2}$ | $\lambda_n \succ \left(\frac{s'_n \log p_n}{n}\right)^{1/2}$ | $s'_n = O(1)$ |

Thank



You