Estimating Covariance Structure in High Dimensions

Ashwini Maurya

Michigan State University
East Lansing, MI, USA

Thesis Director: Dr. Hira L. Koul

Committee Members:
Dr. Yuehua Cui,
Dr. Hyokyoung Hong, and
Dr. Mark A. Iwen

My Research

My research interest consists of two major parts:

Statistical Machine Learning

- Large dimensional covariance and inverse covariance matrices estimation.
- Big data optimization problems.

Quantitative Genetics

- Estimation of genetic heritability for big data.
- Quantitative Trait Loci (QTL) mapping.

Outline

Part I: Covariance Matrix Estimation

- Introduction
- Joint Penalty Method: Optimal Estimation and Theoretical Results
- A Very Fast Algorithm and Simulation Analysis

Part II: Inverse Covariance Matrix Estimation

- A two step approach to Inverse Covariance Matrix Estimation
- Simulations and Analysis of Colon Tumor Gene Expression Data
- A Likelihood Based Approach: Joint Convex Penalty

- Summary and Future Work Directions

Part I

Covariance Matrix Estimation

Introduction

Introduction

Notation:

- Σ_0, Γ_0 : Population covariance and correlation matrices respectively.
- $\Omega_0 = \Sigma_0^{-1}$, $\Psi_0 = \Gamma_0^{-1}$: be their inverse counterparts respectively.
- S, R : Sample covariance and correlation matrix respectively.
- I is the identity matrix of appropriate dimension.

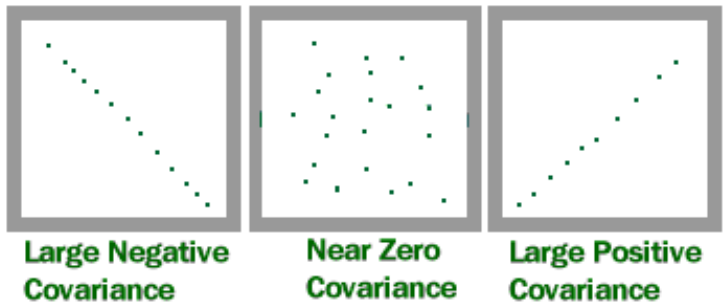
For a matrix M:

- $\|M\|_1$ denote its ℓ_1 norm defined as the sum of absolute values of the entries of matrix M,
- $\|M\|_F$ denote the Frobenius norm of matrix M defined as square root of sum of squared elements of M,
- $\|M\|$ denote the operator norm (also called spectral norm) defined as largest absolute eigenvalue of M,
- M^- denote matrix M where all diagonal elements are set to zero,
- $\sigma_i(M)$ denote the i^{th} largest eigenvalue of M,
- $\bar{\sigma}(M)$ denotes the average of eigenvalues of M, and
- $\|M\|_*$ be its trace norm defined as sum of its singular values.

Introduction: What is a Covariance Matrix?

"Covariance is a measure of how the change in one variable is linearly associated with the change in the other variable."

Figure: Types of Covariance



Sample Covariance Matrix: Let $X := (X_1, X_2, ..., X_n)$ set of p-variate random vectors from a population with mean μ and variance covariance matrix $\mathbf{\Sigma}$. The sample covariance matrix is given by,

$$\mathbf{S} = [[\mathbf{S}_{ij}]] \quad \text{where} \quad \mathbf{S}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(X_{ik} - \bar{X}_i\right)\left(X_{kj} - \bar{X}_j\right).$$

Why We Care About Covariance Matrix?

Introduction: Some Applications

- Statistical Network Analysis
  Network of Neurons: Which part of brain communicates during a given task?
- Climate Data Analysis
  The climate correlations among geographical regions.
- Financial Data Analysis
  Portfolio management in Finance.
- Statistical Genetics
  Genetic networks of quantitative variations in complex traits.
- Many more.

Introduction: 1. Colon Tumor Classification Example

Data:

- Gene expression levels of 62 tissue samples described by 2000 genes.

Two types of tissues:

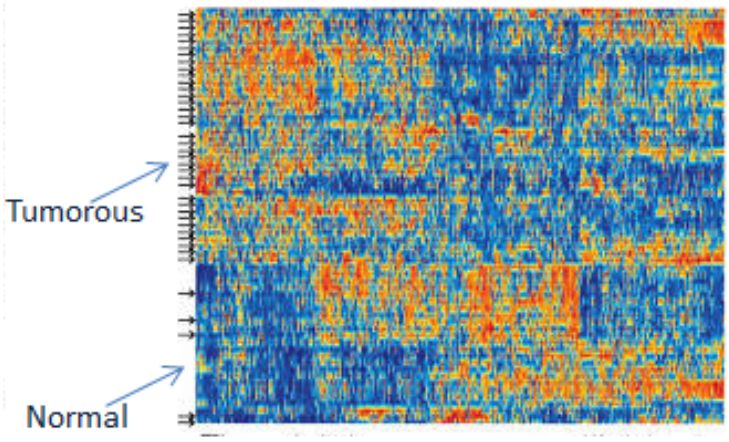
- Normal and Tumorous.

Goal:

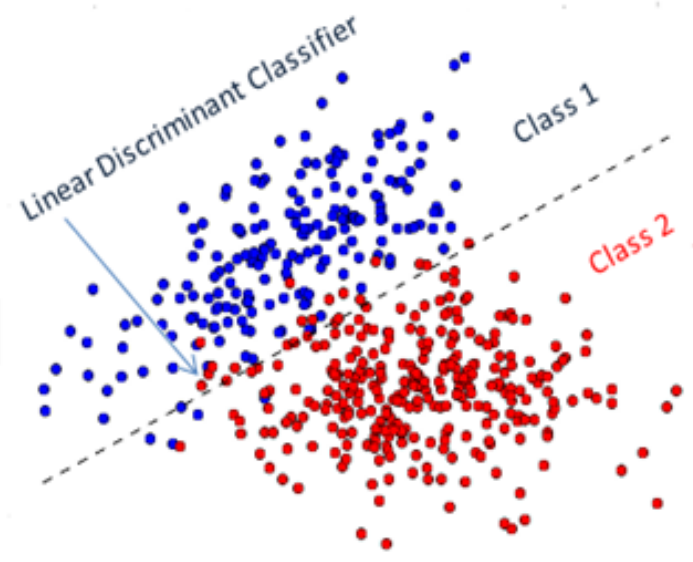
- Classification of normal and tumorous tissues using gene expression data.

How to do it ?

- Using Linear Discriminant Model (LDA)
  - The LDA classifier is function of the inverse covariance matrix.



Source: source: http://genomics-pubs.princeton.edu/oncology/



Introduction: 1. Colon Tumor Classification Example

Data:

- Gene expression levels of 62 tissue samples described by 2000 genes.

Two types of tissues:

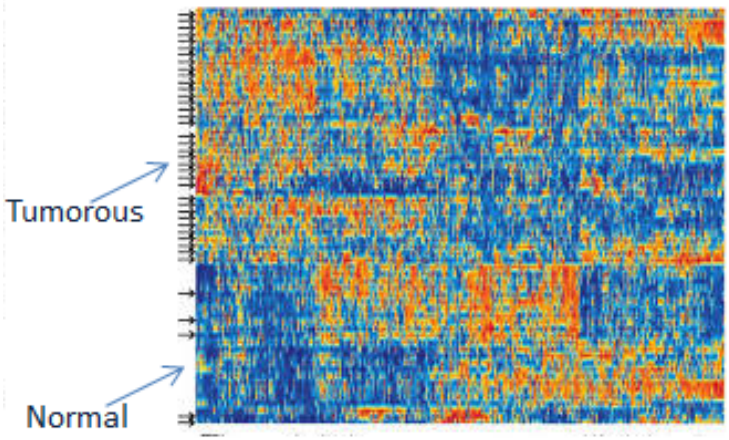
- Normal and Tumorous.

Goal:

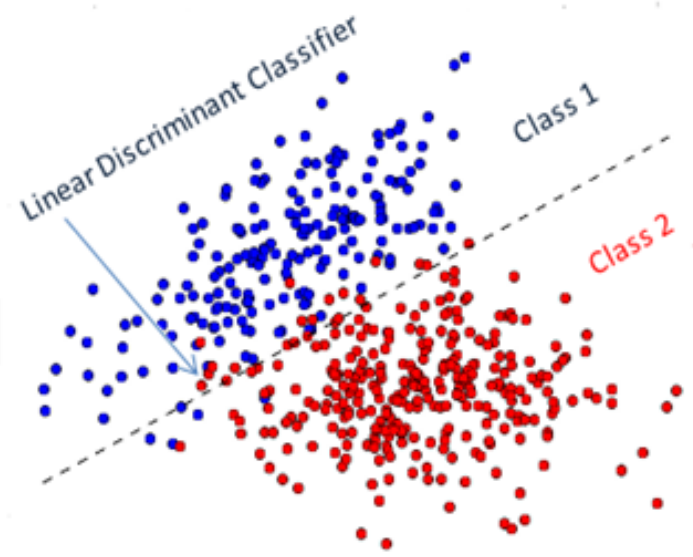
- Classification of normal and tumorous tissues using gene expression data.

How to do it ?

- Using Linear Discriminant Model (LDA)
  - The LDA classifier is function of the inverse covariance matrix.



Source: source: http://genomics-pubs.princeton.edu/oncology/



Introduction: 2. PCA Application in Image Compression

- $\mathbf{X}$ be data matrix of 512×512 gray scale image.
- **Goal:** To compress $\mathbf{X}$ and recover it without loosing much of originality.

- **How to do it ?**
  Representing the $\mathbf{X}$ in lower dimension, which is equivalent to finding set of orthogonal vector $\vec{a}$ such that variance of linear projection $\mathbf{X}\vec{a}$ is maximized.
- It turns out that, $\vec{a}$ is the leading eigenvector of **covariance matrix** $\mathbf{X}^{\mathbf{T}}\mathbf{X}$.

Introduction: 2. PCA Application in Image Compression

- $\mathbf{X}$ be data matrix of 512×512 gray scale image.
- **Goal:** To compress $\mathbf{X}$ and recover it without loosing much of originality.

- **How to do it ?**
  Representing the $\mathbf{X}$ in lower dimension, which is equivalent to finding set of orthogonal vector $\vec{a}$ such that variance of linear projection $\mathbf{X}\vec{a}$ is maximized.
- It turns out that, $\vec{a}$ is the leading eigenvector of **covariance matrix** $\mathbf{X}^{\mathbf{T}}\mathbf{X}$.

Introduction: 2. PCA Application in Image Compression

- $\mathbf{X}$ be data matrix of 512×512 gray scale image.
- **Goal:** To compress $\mathbf{X}$ and recover it without loosing much of originality.

- **How to do it ?**
  Representing the $\mathbf{X}$ in lower dimension, which is equivalent to finding set of orthogonal vector $\vec{a}$ such that variance of linear projection $\mathbf{X}\vec{a}$ is maximized.
- It turns out that, $\vec{a}$ is the leading eigenvector of **covariance matrix** $\mathbf{X^T X}$.

Sample Covariance Matrix Estimator

Low Dimensional Setting: $n < p$, S is a good estimator of its population counterpart. In fact,

- It is consistent. For fixed p, $\mathbb{E}(X^4) < \infty \quad \Rightarrow \quad S \xrightarrow{n\to\infty} \Sigma \quad a.s.$
- It is invertible and extensively used in linear models and time series analysis.
- Its eigenvalues are well behaved and good estimators of their population counterparts.
- It is unbiased.
- It is an approximate Maximum Likelihood Estimator.
- Together $\bar{X}$, the set $(\bar{X}, S)$ is sufficient statistics for the family of Gaussian distributions.

Because of the these properties, it is extensively used for both structure estimation and prediction in many data analysis applications.

What is wrong with Sample Covariance Matrix?

High Dimensional Setting: p is **very large** compared to n.

S is not a good estimator of Σ. In fact,

- For $n < p$, no more positive definite and invertible.
- Its eigenvalues are over-dispersed. Moreover p - n eigenvalues are exactly equal to zero.
- Not Sparse. Generally very noisy and therefore biased for Σ, and
- LDA breaks down if $p/n \to \infty$.

Not very useful in high dimensional setting.

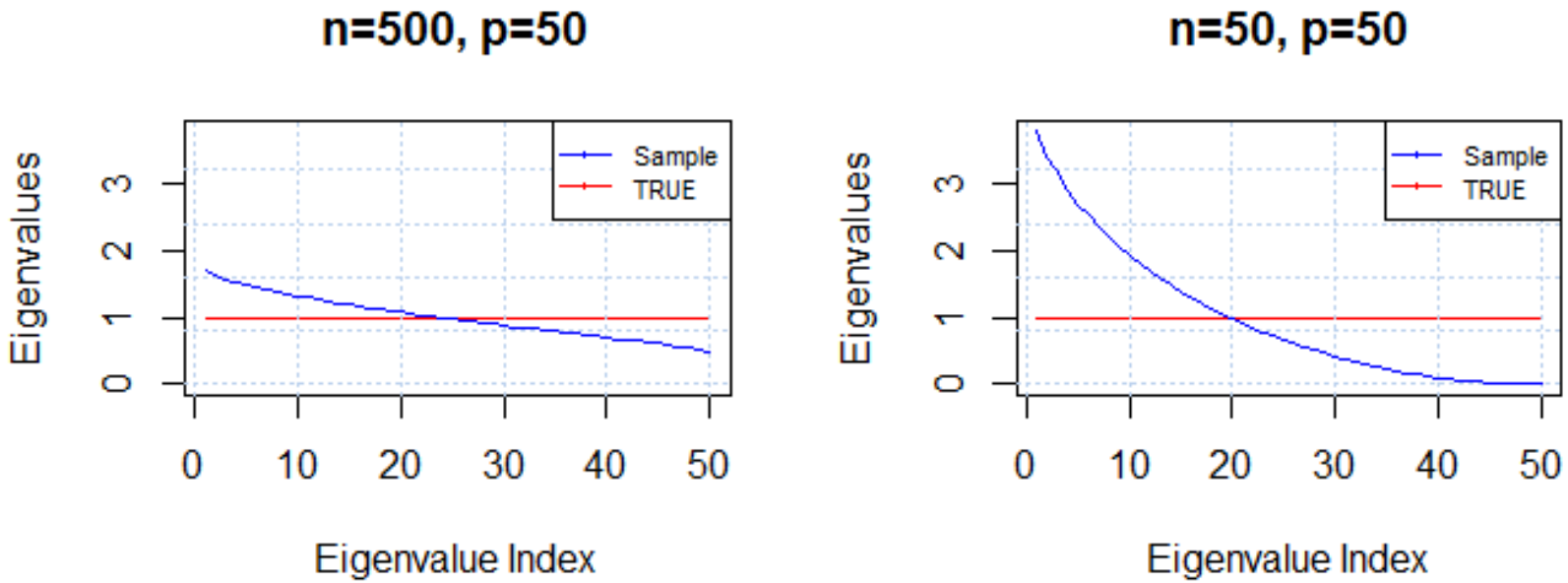
What is Wrong with Sample Covariance Matrix?

1. Over-dispersion in sample eigenvalues.

Data: Multivariate Gaussian with mean vector zero and Identity covariance matrix.

- case (i): n=500,p=50,
- case (ii): n=50,p=50.

Figure: Eigenvalues of Sample and True Covariance Matrices

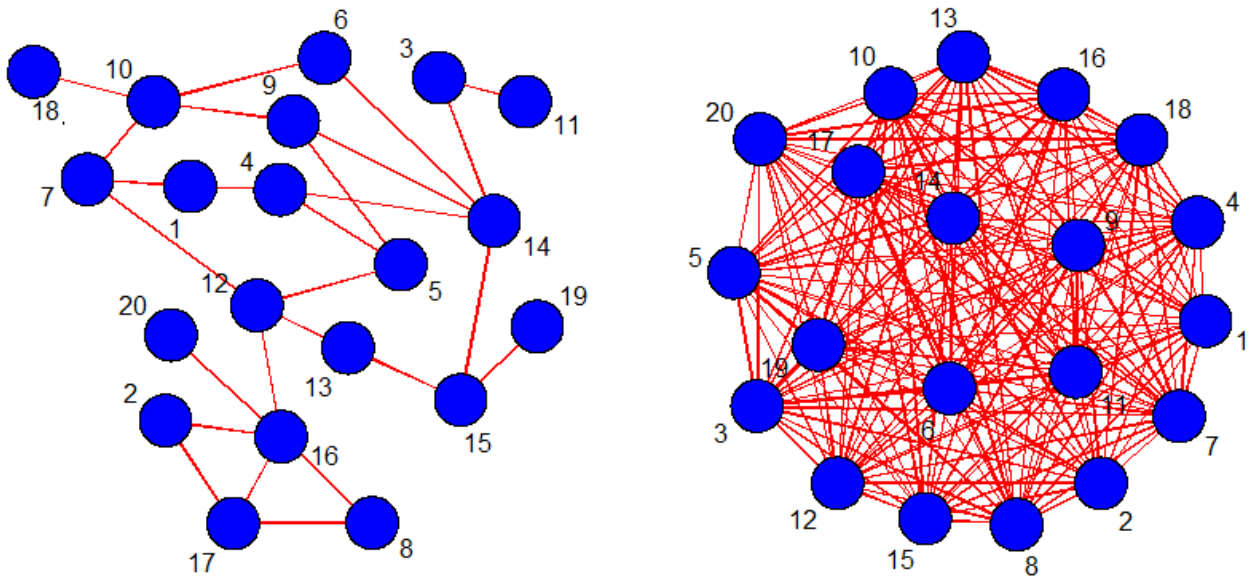


What is Wrong with Sample Covariance Matrix

2. Lack of Sparsity

Data: Multivariate Gaussian with mean vector zero and neighborhood type covariance matrix, n=20,p=20.

Figure: Graph of Population and Sample Covariance Matrix



Notion of Well-Condition in High-Dimensions

Notion of Well-Condition in High-Dimensions

"By a well-conditioned covariance matrix, we mean that its condition number (ratio of maximum and minimum eigenvalues) is bounded from above by a positively finite constant".

Why we need well-conditioned covariance matrix?

1. Invertible and positive definiteness.

   LDA requires inverse of covariance matrix.

2. Well-conditioned estimator reduces estimation error. [Ledoit and Wolf [2004]]

   Regression Analysis: The regression coefficients estimates are function of inverse covariance matrix of independent variables. The estimated coefficients tend to have huge bias if the corresponding inverse covariance matrix is ill-conditioned.

3. Improved estimation of eigenvalues

Notion of Well-Condition in High-Dimensions

"By a well-conditioned covariance matrix, we mean that its condition number (ratio of maximum and minimum eigenvalues) is bounded from above by a positively finite constant".

Why we need well-conditioned covariance matrix?

1. Invertible and positive definiteness.

LDA requires inverse of covariance matrix.

2. Well-conditioned estimator reduces estimation error. [Ledoit and Wolf [2004]]

Regression Analysis: The regression coefficients estimates are function of inverse covariance matrix of independent variables. The estimated coefficients tend to have huge bias if the corresponding inverse covariance matrix is ill-conditioned.

3. Improved estimation of eigenvalues

Notion of Well-Condition... Literature Review

- **Stein's Estimator:** Stein [1975]

  Let $\mathbf{S} = UDU^T$ be the eigen-decomposition of S, where D is diagonal matrix of eigenvalues, and U is matrix of eigenvectors. Let $D =$diag$(d_{11}, d_{22}, \cdots, d_{pp})$. The Stein's estimator is

$$\hat{\Sigma} = UD^{new}U^T, \tag{0.1}$$

  where

$$D^{new} = \text{diag}(d_1^{new}, d_2^{new}, \cdots, d_p^{new}), \qquad \text{with}$$

$$d_{ii}^{new} = nd_{ii} \Big/ \left(n - p + 1 + 2d_{ii} \sum_{i \neq j}^{p} \frac{1}{d_{ii} - d_{jj}} \right).$$

  Advantage: Reduces the over-dispersion in eigenvalues.

  Limitations: Not sparse, not necessarily positive definite, not suitable in high dimensions.

  Related work: Stein [1986],Dey and Srinivasan [1985], Lin and Perlman [1985].

Notion of Well-Condition... Literature Review

- **Ledoit and Wolf's Estimator:** Ledoit and Wolf [2004].

$$\hat{\Sigma} = \rho S + (1-\rho) I, \qquad \rho \text{ is estimated from data.} \tag{0.2}$$

Advantage: Well-Conditioned.
Limitation: Uniform shrinkage, not sparse.

- Won et al. [2012]

$$\hat{\Sigma} = \arg\max_{\Sigma} \quad L(S, \Sigma) \quad \text{subject to} \quad \frac{\sigma_{max}(\Sigma)}{\sigma_{max}(\Sigma)} \leq \kappa_{max}. \tag{0.3}$$

$\hat{\Sigma}$ invertible if κ_{max} finite and well-conditioned if κ_{max} is moderate.

Advantage: Well-conditioned
Limitation: Hard to say if the it gives improved estimation of eigen-structure.

Notion of Sparsity in High-Dimensions

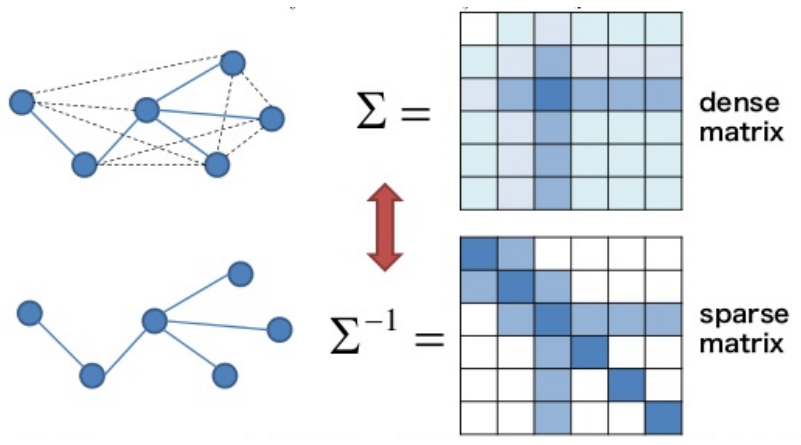
Notion of Sparsity in High Dimensions

" A covariance matrix is said to be sparse if most of its entries are zero. Equivalently most of the variables are uncorrelated with each other."

Why we need a sparse estimation of a covariance matrix?

- **To avoid curse of dimensionality:**
  For a $p \times p$ matrix, total # of free parameters $= p(p+1)/2$. When $n < p$, this is ill defined problem.
- The high dimensional covariance structures can be represented by few parameters, which is the case in many scientific studies.

Figure: Dense and Sparse Covariance and Precision Matrices



Notion of Sparsity in High Dimensions

Real Life Data Examples:

- **Correlation network of colon tumor gene expression data:** Study [Alon et al. [1999]] shows that among given set of more than 2000 gene expression, each gene shows a strong correlation with on the order of 1% of other genes.

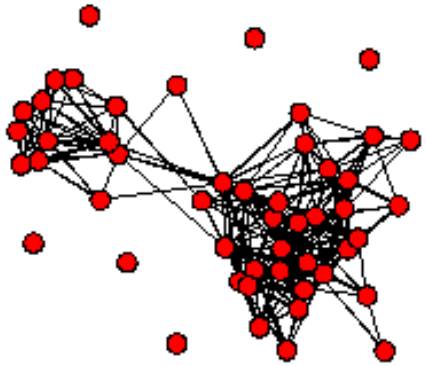


Figure: Correlation network of colon tumor gene expression data

- **Climate data analysis:** Study of correlation of weather temperature across a geographical area.

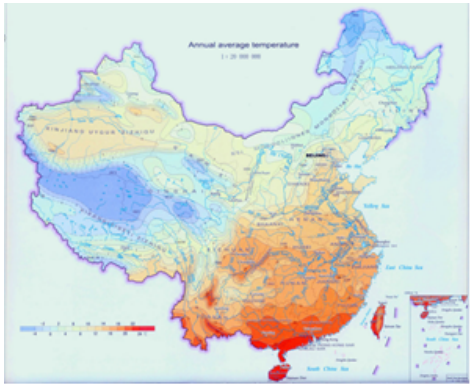


Figure: Weather temperature across China, 07/22/2011

Notion of Sparsity in High Dimensional: Literature Review

Two broad class of sparse covariance matrices:

1. Natural ordering among variables:

The variables far apart are weakly correlated.

Example: Time series analysis.

Estimation procedure:

Assumes structure such as underlying covariance matrix is Toeplitz type.

Bickel and Levina [2008a,b], Cai et al. [2015]

Figure: Toeplitz type matrix

$$\begin{pmatrix} 2 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

Notion of Sparsity in High Dimensional: Literature Review

2. No natural ordering among variables:

A prior knowledge of any canonical ordering among variables is not available.

Example: Gene expression data.

Earlier work:

(i) Bien and Tibshirani [2011]

$$\hat{\Sigma} = \underset{\Sigma \succ 0}{\operatorname{argmin}} \Big[\log(\det(\Sigma)) + \mathsf{tr}(S\Sigma^{-1}) + \lambda \|\Sigma\|_1 \Big], \quad \lambda > 0.$$

(ii) Positive Definite Sparse Covariance Estimator (PDSCE):[Rothman [2012]]

$$\hat{\Sigma} = \underset{\Sigma = \Sigma^T}{\operatorname{argmin}} \Big[\|\Sigma - S\|_F^2 + \lambda \|\Sigma^-\|_1 - \gamma \log(\det(\Sigma)) \Big], \quad \lambda, \gamma > 0.$$

(iii) Chaudhuri et al. [2007], Xue et al. [2012].

Simultaneous Estimation of Sparse and Well-Conditioned Covariance Matrices

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

Joint Penalty (JPEN): A Well-Conditioned and Sparse Estimation

Goal: A Sparse and Well-Conditioned Estimator. (First consider estimation of correlation matrix Γ.)

Let $\hat{\Gamma}$ be the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right]. \quad (0.4)$$

- For a large value of λ, the penalty $\lambda\|\Gamma^-\|_1$ shrinks the smaller elements of Γ to zero.
- By $tr(\Gamma) = tr(R)$, the total variation in $\hat{\Gamma}$ remains as that in R.
- For a large value of γ, the variance of eigenvalue penalty, reduces the over-dispersion in the covariance matrix.

It turns out that $\hat{\Gamma}$ is sparse and reduces the over-dispersion in eigenvalues of R. However simulation shows that $\hat{\Gamma}$ need not be positive definite for all values of (λ, γ).

JPEN: A Well-Conditioned and Sparse Estimation

Proposed Estimator:

The proposed JPEN estimator is given by:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T | (\lambda,\gamma)\in\hat{\mathbb{S}}_1^R, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right].$$

where

$$\hat{\mathbb{S}}_1^R = \left\{ (\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\frac{log\ p}{n}} : \lambda < \frac{\sigma_{min}(R+\gamma I)}{C\ \sigma_{max}(sign(R))} \right\},$$

$C \geq 0.5$, and $sign(R)$ is matrix of signs of elements of R.

The corresponding JPEN covariance matrix estimator is:

$$\hat{\Sigma}_R = D\hat{\Gamma}D^T$$

where D is diagonal matrix of sample standard deviations.

JPEN: A Well-Conditioned and Sparse Estimation

Proposed Estimator:

The proposed JPEN estimator is given by:

$$\hat{\Gamma} = \underset{\Gamma=\Gamma^T \,|\, (\lambda,\gamma)\in\hat{\mathbb{S}}_1^R, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right].$$

where

$$\hat{\mathbb{S}}_1^R = \left\{ (\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\frac{log\ p}{n}} : \lambda < \frac{\sigma_{min}(R+\gamma I)}{C\ \sigma_{max}(sign(R))} \right\},$$

$C \geq 0.5$, and $sign(R)$ is matrix of signs of elements of R.

The corresponding JPEN covariance matrix estimator is:

$$\boxed{\hat{\Sigma}_R = D\hat{\Gamma}D^T}$$

where D is diagonal matrix of sample standard deviations.

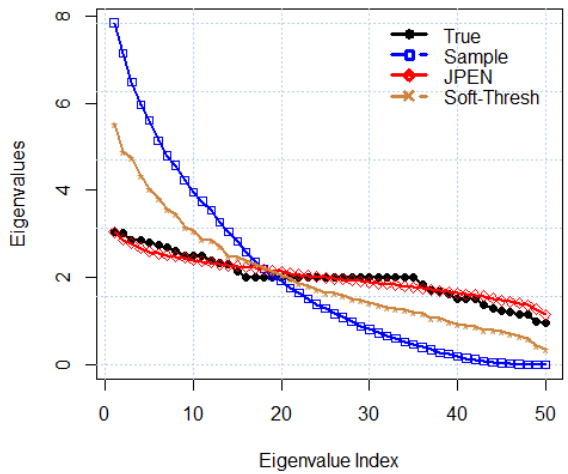
JPEN: Why we need two penalties?

Case (i): For $\gamma = 0$, the JPEN estimator is soft-thresholding estimator and it is given by:

$$\hat{\Sigma}_{ii} = s_{ii}, \quad \text{and} \quad \hat{\Sigma}_{ij} = \text{sign}(s_{ij}) \max\left(|s_{ij}| - \frac{\lambda}{2}, 0\right), \qquad i \neq j. \quad (0.5)$$

- A sufficiently large value of λ will result in sparse covariance matrix estimate.
- However the estimator $\hat{\Sigma}$ of (0.2) need not be positive definite.
- Moreover it is hard to say whether it overcomes the over-dispersion in the sample eigenvalues.
- Eigenvalues of the JPEN estimator are well aligned with those of the true covariance matrix.

Figure: Comparison of Eigenvalues of Covariance Matrix Estimates



JPEN: Why we need two penalties?

Case (ii): For $\lambda = 0$, the JPEN estimator is given by:

$$\hat{\Sigma} = (S + \gamma I)/(1 + \gamma). \tag{0.6}$$

Note that,

$$\sigma_{min}(\hat{\Sigma}) = (\sigma_{min}(S) + \gamma)/(1 + \gamma) \geq \gamma/(1 + \gamma) > \epsilon,$$

for all $\epsilon > c/(1 - c)$.

Therefore the variance of eigenvalues penalty improves S to be well-conditioned.

JPEN: A More Generic Estimator

For any $\{a_i : 0 < a_i < \infty, i = 1, 2, \cdots, p\}$; a weighted JPEN correlation matrix estimator is given by:

$$\hat{\Gamma}^a = \underset{\Gamma=\Gamma^T | (\lambda,\gamma)\in\hat{\mathbb{S}}_1^{R,a}, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} a_i \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right],$$

where

$$\hat{\mathbb{S}}_1^{R,a} = \left\{ (\lambda, \gamma) : \lambda, \gamma > 0 : \lambda < \frac{2\sigma_{min}(R)(1 + \gamma\ max(a_i)^{-1})}{(1 + \gamma\ min(a_i))^{-1} p} + \frac{\gamma\ min(a_i)}{p} \right\}.$$

The covariance matrix estimator is:

$$\boxed{\hat{\Sigma} = D\hat{\Gamma}^a D^T.}$$

Advantage:

- Choice of weights a_i allows non-uniform shrinkage of eigenvalues towards their mean.
- The weighted estimator yields improved eigenvalues estimation.

JPEN: A More Generic Estimator

For any $\{a_i : 0 < a_i < \infty, i = 1, 2, \cdots, p\}$; a weighted JPEN correlation matrix estimator is given by:

$$\hat{\Gamma}^a = \underset{\Gamma=\Gamma^T | (\lambda,\gamma)\in\hat{\mathbb{S}}_1^{R,a}, tr(\Gamma)=tr(R)}{\operatorname{argmin}} \left[||\Gamma - R||_F^2 + \lambda \|\Gamma^-\|_1 + \gamma \sum_{i=1}^{p} a_i \{\sigma_i(\Gamma) - \bar{\sigma}(\Gamma)\}^2 \right],$$

where

$$\hat{\mathbb{S}}_1^{R,a} = \left\{ (\lambda, \gamma) : \lambda, \gamma > 0 : \lambda < \frac{2\sigma_{min}(R)(1 + \gamma\ max(a_i)^{-1})}{(1 + \gamma\ min(a_i))^{-1} p} + \frac{\gamma\ min(a_i)}{p} \right\}.$$

The covariance matrix estimator is:

$$\boxed{\hat{\Sigma} = D\hat{\Gamma}^a D^T.}$$

Advantage:

- Choice of weights a_i allows non-uniform shrinkage of eigenvalues towards their mean.
- The weighted estimator yields improved eigenvalues estimation.

JPEN: A Very Fast Algorithm

The JPEN correlation matrix estimator is:

$$\hat{\Gamma}_{ij} = \frac{1}{1+\gamma}\ sign(R_{ij}) * max\{abs(R_{ij} + \gamma\ I) - \frac{\lambda}{2}, 0\},\ i \neq j \text{ and;}$$
$$\hat{\Gamma}_{ii} = 1.$$

Choice of λ and γ:
For given value of γ, we can find the value of λ satisfying

$$\lambda < \frac{\sigma_{min}(R+\gamma I)}{C\ \sigma_{max}(sign(R))}.$$

For $C \geq 0.5$, the estimator is positive definite. A smaller value of C yields a solution which is more sparse but the estimator may not remain positive definite.

JPEN: A Very Fast Algorithm

The JPEN correlation matrix estimator is:

$$\hat{\Gamma}_{ij} = \frac{1}{1+\gamma}\ sign(R_{ij}) * max\{abs(R_{ij} + \gamma\ I) - \frac{\lambda}{2}, 0\},\ i \neq j \text{ and;}$$
$$\hat{\Gamma}_{ii} = 1.$$

Choice of λ and γ:
For given value of γ, we can find the value of λ satisfying

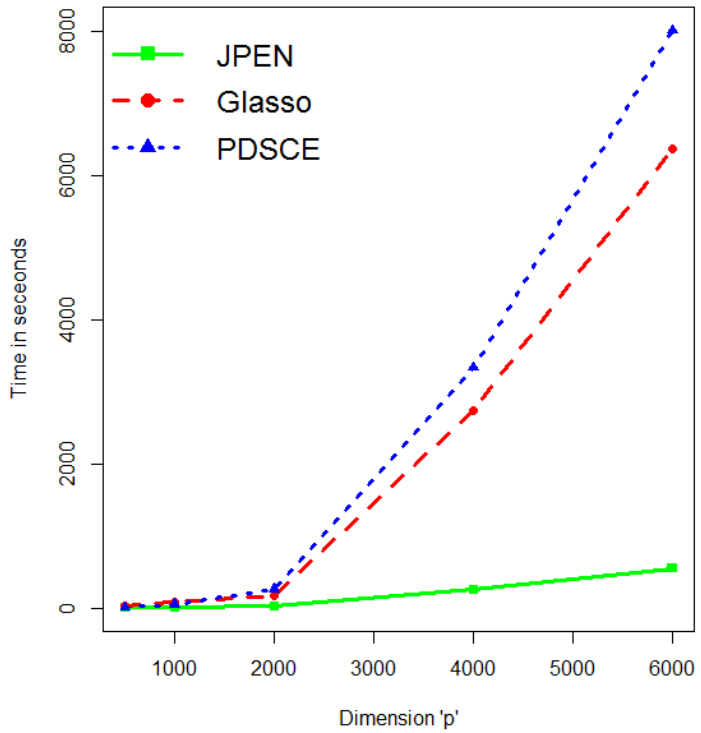
$$\lambda < \frac{\sigma_{min}(R+\gamma I)}{C\ \sigma_{max}(sign(R))}.$$

For $C \geq 0.5$, the estimator is positive definite. A smaller value of C yields a solution which is more sparse but the estimator may not remain positive definite.

JPEN: Computational Complexity

- **JPEN:** $O(p^2)$ (very fast) as there are $(p^2 + 2p)$ multiplication, and at most p^2 operations for entry-wise maximum computations.
- **Graphical Lasso** (Glasso) [Friedman et al. [2007]], **PDSCE** [Rothman, A. [2012]]: At least $O(p^3)$, very slow for large p.
- Here $n = 500$, $p = 500, 1000, 2000, 4000, 6000$.
- Total time includes computation of optimal tuning parameters.

Figure: Timing comparison of JPEN, Graphical Lasso, PDSCE.



Asymptotic Consistency of JPEN Estimators

JPEN: Asymptotic Consistency of Estimators- Set up

Assumptions:

- A0: $X_1, X_2, \cdots, X_n$ be mean zero, sub-Gaussian random vectors with true covariance matrix Σ_0.
- A1: With $E = \{(i,j) : i \neq j, \Sigma_{0ij} \neq 0\}$, the $|E| \leq s$ for some positive integer s.
- A2: There exists some finite positive real number $\bar{k} > 0$ such that $1/\bar{k} \leq \sigma_{min}(\Sigma_0) \leq \sigma_{max}(\Sigma_0) \leq \bar{k}$.

Goal:

- Establish theoretical consistency of estimators in both operator and Frobenius norm.
- The applicability of proposed methods in high dimensional set up.

JPEN: Asymptotic Consistency of Estimators- Results

Under the assumptions A0,A1,A2:

Correlation Matrix Estimation

$$\boxed{\|\hat{\Gamma} - \Gamma_0\|_F = O_P\left(\sqrt{\frac{s\ log\ p}{n}}\right).}$$

Covariance Matrix Estimation

$$\boxed{\|\hat{\Sigma}_R - \Sigma_0\| = O_P\left(\sqrt{\frac{(s+1)log\ p}{n}}\right).}$$

- The JPEN estimator $\hat{\Sigma}_R$ is mini-max optimal in operator norm.
- In high dimensional setting, for sparse matrices with $s = O(log\ p)$, the JPEN estimator is consistent in operator norm even when the dimension grows exponentially with sample size.

JPEN: Asymptotic Consistency of Estimators- Results

Under the assumptions A0,A1,A2:

Correlation Matrix Estimation

$$\boxed{\|\hat{\Gamma} - \Gamma_0\|_F = O_P\left(\sqrt{\frac{s\ log\ p}{n}}\right).}$$

Covariance Matrix Estimation

$$\boxed{\|\hat{\Sigma}_R - \Sigma_0\| = O_P\left(\sqrt{\frac{(s+1)log\ p}{n}}\right).}$$

- The JPEN estimator $\hat{\Sigma}_R$ is mini-max optimal in operator norm.
- In high dimensional setting, for sparse matrices with $s = O(log\ p)$, the JPEN estimator is consistent in operator norm even when the dimension grows exponentially with sample size.

Joint Penalty: Inverse Covariance Matrix Estimation

JPEN: Inverse Covariance Matrix Estimation

Computing $\hat{\Omega}$ is a two step approach:

- First compute a well-conditioned estimator $\hat{\Gamma}$ of Γ.
- Use $\hat{\Gamma}^{-1}$ as starting point to estimate the JPEN inverse correlation matrix estimator.

The JPEN estimator of inverse correlation matrix Γ_0^{-1} is given by

$$\hat{\Psi}^{-1} = \underset{\Psi=\Psi^T|(\lambda,\gamma)\in\hat{\mathbb{S}}_2^R, tr(\Psi)=tr(\hat{\Gamma}^{-1})}{\operatorname{argmin}} \left[||\Psi - \hat{\Gamma}^{-1}||_F^2 + \lambda \|\Psi^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Psi) - \bar{\sigma}(\Psi)\}^2 \right],$$

where

$$\hat{\mathbb{S}}_2^R = \left\{ (\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\frac{log\ p}{n}} : \lambda < \frac{\sigma_{min}(\hat{\Gamma}^{-1} + \gamma I)}{C_1\ \sigma_{max}(sign(\hat{\Gamma}^{-1}))} \right\}$$

and $C_1 \geq 0.5$.

JPEN: Inverse Covariance Matrix Estimation

An estimator of inverse covariance matrix Ω is given by

$$\boxed{\hat{\Omega}_R = D^{-1}\hat{\Psi}D^{-1^T}}$$

where D is diagonal matrix of sample standard deviations.

JPEN: Asymptotic Consistency of Estimators: Set up

Assumptions:

- B0: $X_1, X_2, \cdots, X_n$ be mean zero, sub-Gaussian random vectors with true covariance matrix Σ_0. Denote $\Omega_0 := \Sigma_0^{-1}$ as the true inverse covariance matrix.
- B1: With $E = \{(i,j) : i \neq j, \Omega_{0ij} \neq 0\}$, the $|E| \leq s$ for some positive integer s.
- B2: There exists some finite positive real number $\bar{k} > 0$ such that $1/\bar{k} \leq \sigma_{min}(\Omega_0) \leq \sigma_{max}(\Omega_0) \leq \bar{k}$.

Goal:

- Establish theoretical consistency of estimators in both operator and Frobenius norm.
- The applicability of proposed methods in high dimensional set up.

JPEN: Asymptotic Consistency of Estimators: Results

Under the assumptions B0,B1,B2:

Inverse Correlation Matrix Estimation:

$$\boxed{\|\hat{\Psi} - \Psi_0\|_F = O_P\left(\sqrt{\frac{s\ log\ p}{n}}\right)}$$

Inverse Covariance Matrix Estimation:

$$\boxed{\|\hat{\Omega}_R - \Omega_0\| = O_P\left(\sqrt{\frac{(s+1)log\ p}{n}}\right).}$$

- The JPEN estimator $\hat{\Omega}_R$ is mini-max optimal in operator norm.
- In high dimensional setting, for sparse matrices with $s = O(log\ p)$, the JPEN estimator is consistent in operator norm even when the dimension grows exponentially with sample size.

JPEN: Asymptotic Consistency of Estimators: Results

Under the assumptions B0,B1,B2:

Inverse Correlation Matrix Estimation:

$$\boxed{\|\hat{\Psi} - \Psi_0\|_F = O_P\left(\sqrt{\frac{s\ log\ p}{n}}\right)}$$

Inverse Covariance Matrix Estimation:

$$\boxed{\|\hat{\Omega}_R - \Omega_0\| = O_P\left(\sqrt{\frac{(s+1)log\ p}{n}}\right).}$$

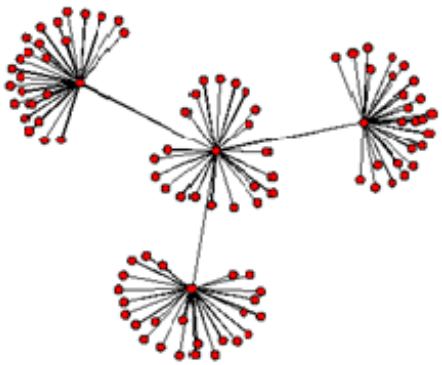
- The JPEN estimator $\hat{\Omega}_R$ is mini-max optimal in operator norm.
- In high dimensional setting, for sparse matrices with $s = O(log\ p)$, the JPEN estimator is consistent in operator norm even when the dimension grows exponentially with sample size.

Simulation Studies and an Application to Colon Tumor Gene Expression Data

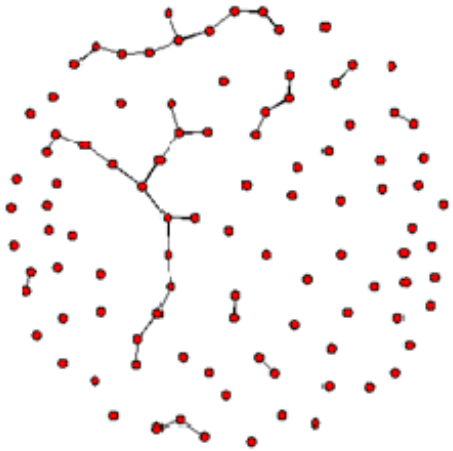
Simulated Examples: Settings

Model: $Y_i \sim MVN(\mathbf{0}, \Sigma_0)$. $n = 100$ and $p = 500, 1000$. We assume following structures of Σ_0.

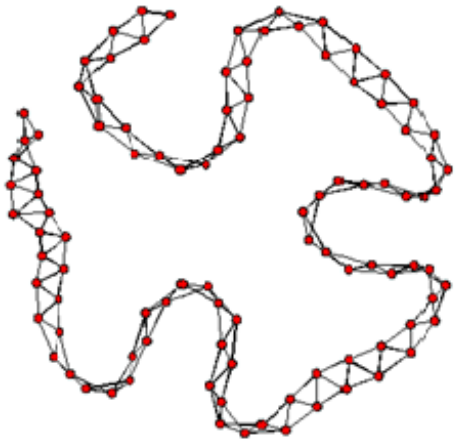
Hub



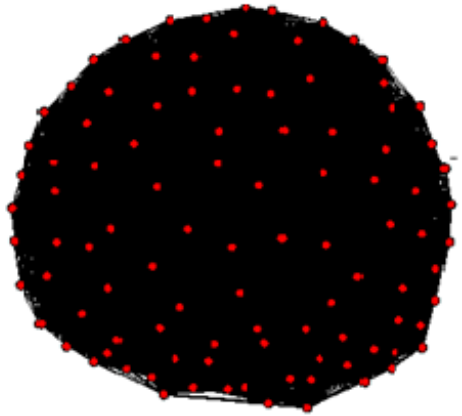
Neighborhood



Toeplitz



Dense



Simulated Examples: Settings

Competing Methods:

- Covariance Matrix Estimation:
  - JPEN: Joint Penalty,
  - Glasso: Graphical Lasso [Friedman et al. (2007)],
  - Ledoit-Wolf Estimator of Covariance Matrix [Ledoit and Wolf (2004)],
  - PDSCE: Positive Definite Sparse Covariance Matrix Estimator,
  - BLThresh: Bickel and Levina's Thresholding Estimator [Bickel and Levina (2008)].
- Inverse Covariance Matrix Estimation
  - JPEN,
  - Graphical Lasso (Glasso),
  - SPICE: Sparse Permutation Invariant Covariance Estimation [Rothman et al. (2008)].

Performance Criteria: Average Relative Error (ARE)

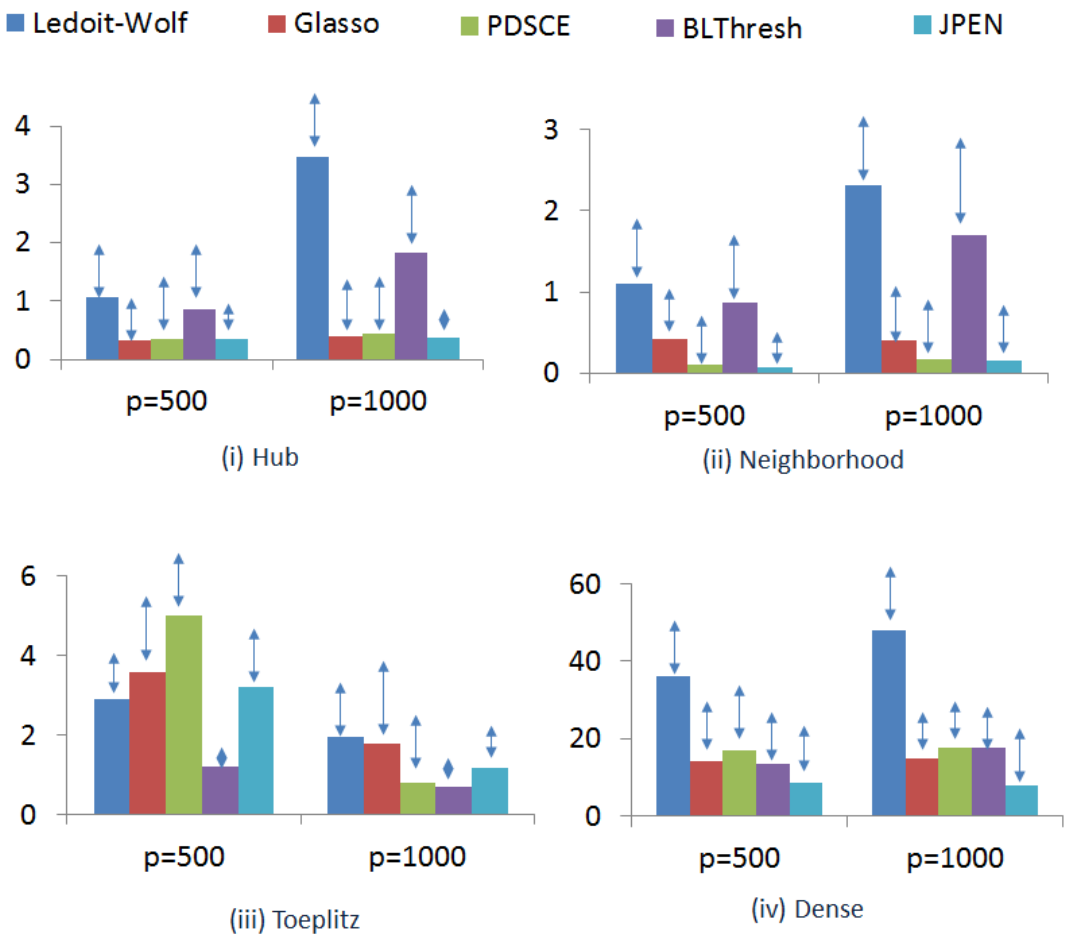
$$ARE(\Sigma_0, \hat{\Sigma}) = \frac{|log(L(S, \hat{\Sigma})) \; - \; log(L(S, \Sigma_0))|}{|log(L(S, \Sigma_0))|},$$

where $L(S, \cdot)$ is likelihood function of multivariate normal distribution.

Simulated Examples: Results

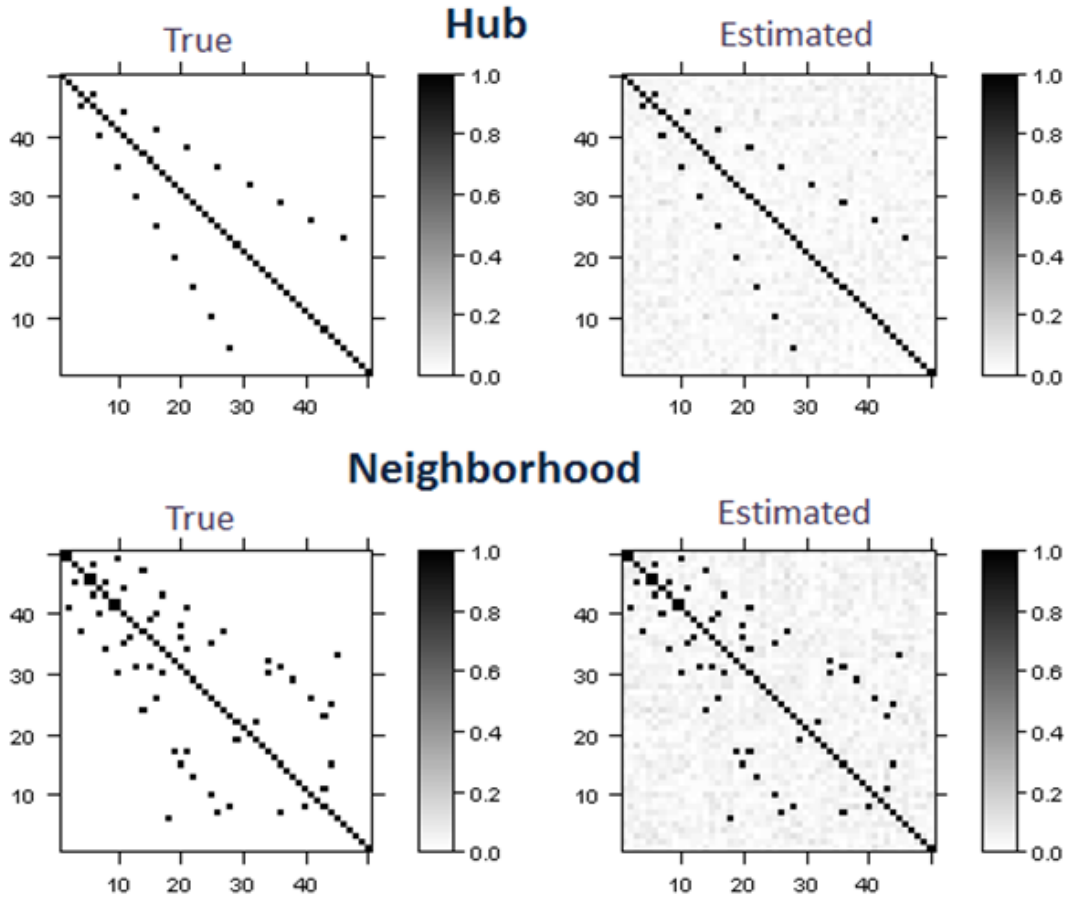
- The height of a bar corresponds to ARE. Smaller bars are better.
- The size of arrow corresponds to standard error. Smaller arrows are better.

Figure: Average relative error and standard errors based on 100 replications



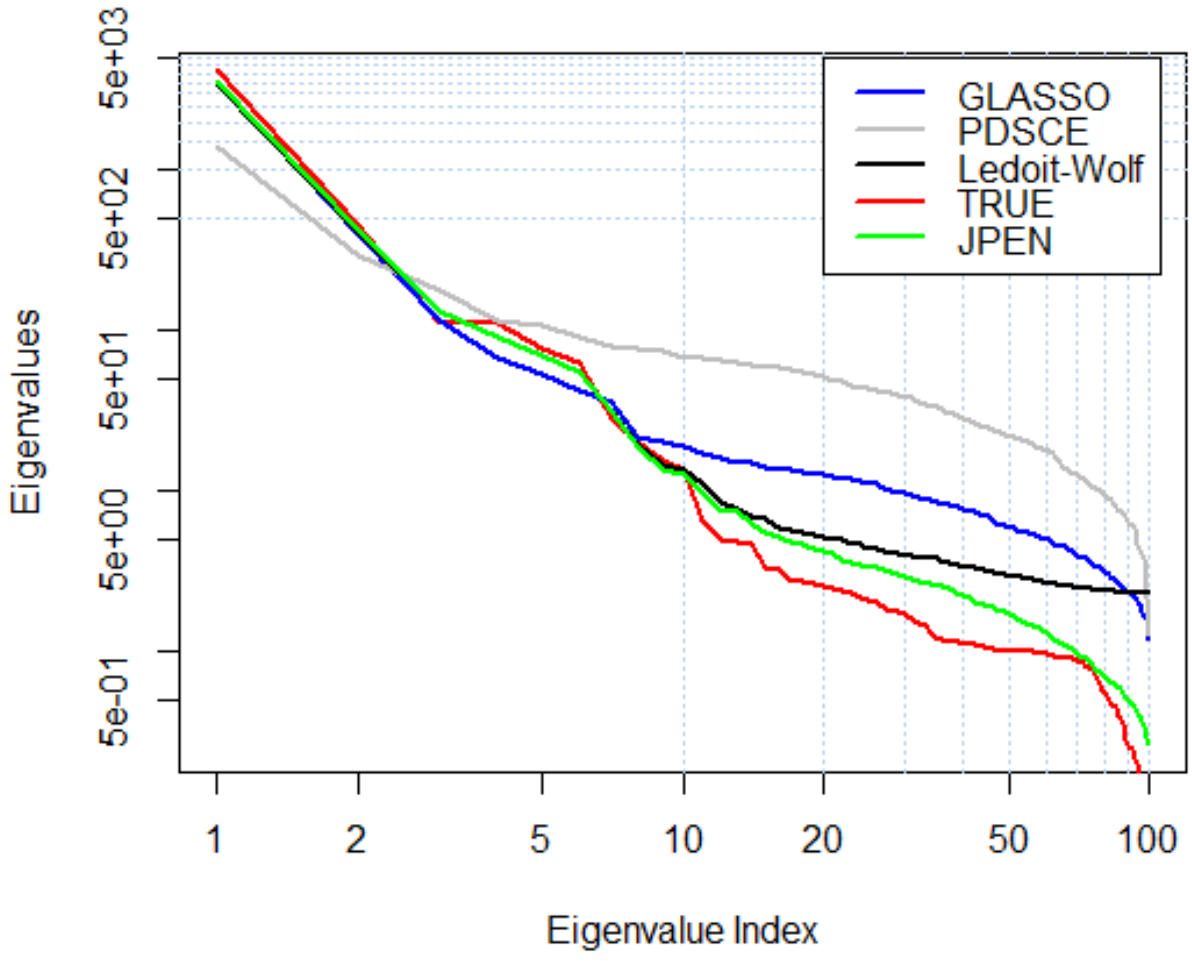
Simulated Example : Sparse (Zeros) Recovery

Figure: Heat-map of zeros identified in covariance matrix out of 50 realizations. White color is 50/50 zeros identified, black color is 0/50 zeros identified.



Simulated Example : Eigenvalues Recovery

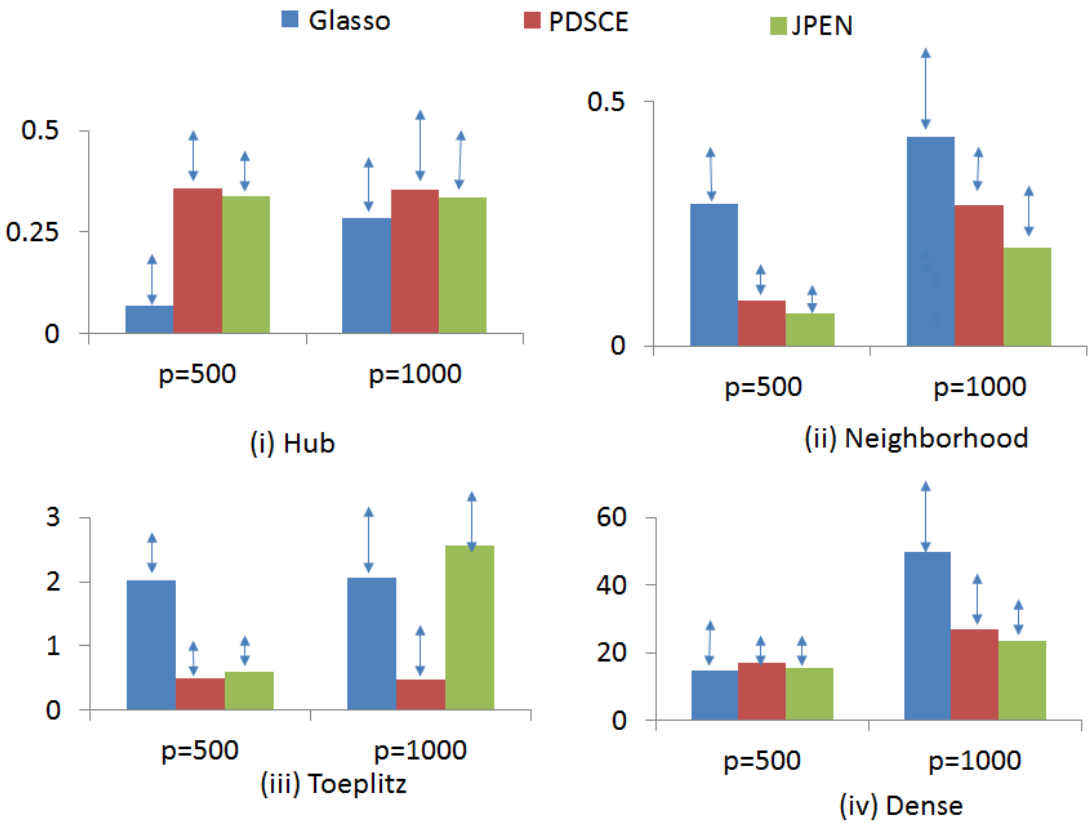
Figure: Recovery of population eigenvalues of Dense type of covariance matrix. Eigenvalues vary from 0.005 to 3872.



Inverse Covariance Matrix Estimation: Simulated Results

- JPEN performs better in neighborhood and dense setting.

Figure: Average relative error and standard errors



Application to Colon Tumor Gene Expression Data

Data: Colon adenocarcinoma tissue samples (40 tumorous and 22 normal) described by 2000 genes.

Performance Comparison:

- Covariance matrix based methods: Graphical Lasso, SPICE, JPEN.

  LDA Rule: Classify an observation x to either class k using ,

  $$\delta_k(x) = \arg\max_k \left\{ x^T \hat{\Omega} \hat{\mu}_k \; - \; \frac{1}{2} \hat{\mu}_k \hat{\Omega} \hat{\mu}_k \; + \; log(\hat{\pi}_k) \right\},$$

  where $\hat{\pi}_k$ and $\hat{\mu}_k$ are the proportion and sample mean of class k observations, and $\hat{\Omega}$ is an estimate of the inverse of the common covariance matrix.

- Other traditional methods: Logistic Regression, Support Vector Machines, Naive Bayes.

Application to Colon Tumor Gene Expression Data

- Further reduce the gene expression set to $p = 50, 100, 200$ genes based on ℓ_1 regularized logistic regression.
- Tuning parameters are selected based on 5-fold cross validation error.

Table: *Mean and standard errors of classification errors based on 100 replicates in %.*

Method	p=50	p=100	p=200
Logistic Regression	21.0(0.84)	19.31(0.89)	21.5(0.85)
SVM	16.70(0.85)	16.76(0.97)	18.18(0.96)
Naive Bayes	13.3(0.75)	14.33(0.85)	14.63(0.75)
Glasoo	10.9(1.3)	9.4(0.89)	9.8(0.90)
SPICE	**9.0(0.57)**	9.1(0.58)	10.2(0.52)
JPEN	9.9(0.98)	**8.9(0.93)**	**8.2(0.81)**

- Covariance matrix based methods outperform other methods.
- With full data analysis of 2000 genes, the classification error was 32%.

Application to Colon Tumor Gene Expression Data

- Further reduce the gene expression set to $p = 50, 100, 200$ genes based on ℓ_1 regularized logistic regression.
- Tuning parameters are selected based on 5-fold cross validation error.

Table: *Mean and standard errors of classification errors based on 100 replicates in %.*

Method	p=50	p=100	p=200
Logistic Regression	21.0(0.84)	19.31(0.89)	21.5(0.85)
SVM	16.70(0.85)	16.76(0.97)	18.18(0.96)
Naive Bayes	13.3(0.75)	14.33(0.85)	14.63(0.75)
Glasoo	10.9(1.3)	9.4(0.89)	9.8(0.90)
SPICE	**9.0(0.57)**	9.1(0.58)	10.2(0.52)
JPEN	9.9(0.98)	**8.9(0.93)**	**8.2(0.81)**

- Covariance matrix based methods outperform other methods.
- With full data analysis of 2000 genes, the classification error was 32%.

Summary

- **Broadly Applicable New Method:** JPEN covariance matrix estimation does not assume any particular structure on the data distribution and hence fully non-parametric. It is **broadly** applicable method for any sparse covariance and inverse covariance matrix estimation.
- **Mini-max Optimal:** JPEN is mini-max optimal in operator norm and hence we expect that PCA will be one of the most important applications of the method. The estimator is also consistent in Frobenius norm.
- **A Very Fast Algorithm:** The proposed algorithm is exact, very fast, and easily scalable to large data analysis problem. The computational complexity is only $O(p^2)$ as compared to $O(p^3)$ of other methods of estimation.
- JPEN estimation allows one to take advantage of any prior structure if known on the eigenvalues of true covariance matrix.

Summary

- **Broadly Applicable New Method:** JPEN covariance matrix estimation does not assume any particular structure on the data distribution and hence fully non-parametric. It is **broadly** applicable method for any sparse covariance and inverse covariance matrix estimation.

- **Mini-max Optimal:** JPEN is mini-max optimal in operator norm and hence we expect that PCA will be one of the most important applications of the method. The estimator is also consistent in Frobenius norm.

- **A Very Fast Algorithm:** The proposed algorithm is exact, very fast, and easily scalable to large data analysis problem. The computational complexity is only $O(p^2)$ as compared to $O(p^3)$ of other methods of estimation.

- JPEN estimation allows one to take advantage of any prior structure if known on the eigenvalues of true covariance matrix.

Summary

- **Broadly Applicable New Method:** JPEN covariance matrix estimation does not assume any particular structure on the data distribution and hence fully non-parametric. It is **broadly** applicable method for any sparse covariance and inverse covariance matrix estimation.

- **Mini-max Optimal:** JPEN is mini-max optimal in operator norm and hence we expect that PCA will be one of the most important applications of the method. The estimator is also consistent in Frobenius norm.

- **A Very Fast Algorithm:** The proposed algorithm is exact, very fast, and easily scalable to large data analysis problem. The computational complexity is only $O(p^2)$ as compared to $O(p^3)$ of other methods of estimation.

- JPEN estimation allows one to take advantage of any prior structure if known on the eigenvalues of true covariance matrix.

Summary

- **Broadly Applicable New Method:** JPEN covariance matrix estimation does not assume any particular structure on the data distribution and hence fully non-parametric. It is **broadly** applicable method for any sparse covariance and inverse covariance matrix estimation.

- **Mini-max Optimal:** JPEN is mini-max optimal in operator norm and hence we expect that PCA will be one of the most important applications of the method. The estimator is also consistent in Frobenius norm.

- **A Very Fast Algorithm:** The proposed algorithm is exact, very fast, and easily scalable to large data analysis problem. The computational complexity is only $O(p^2)$ as compared to $O(p^3)$ of other methods of estimation.

- JPEN estimation allows one to take advantage of any prior structure if known on the eigenvalues of true covariance matrix.

Joint Convex Penalty (JCP): A Likelihood Based Method for Inverse Covariance Matrix Estimation

JCP Inverse Covariance Matrix Estimation

Let $X \sim N_p(0, \Sigma)$, $\Sigma \succ 0$.

Proposed Estimator: JCP inverse covariance matrix estimator is given by

$$\arg\min_{\Omega \succ 0} \quad F(\Omega) := f(\Omega) + g_1(\Omega) + g_2(\Omega), \tag{0.7}$$

where

$$\begin{aligned} f(\Omega) &= -\log(det(\Omega)) + tr(S\Omega), \\ g_1(\Omega) &= \lambda\|\Omega\|_1, \quad \text{and} \quad g_2(\Omega) = \tau\|\Omega\|_*, \quad \lambda, \tau > 0. \end{aligned}$$

- $f(\Omega)$ is a convex function, ℓ_1 norm is a smooth convex function except at origin and trace norm is convex surrogate of rank over the unit ball of spectral norm Fazal [2002]. Therefore the optimization problem above is convex optimization problem with non-smooth constratints.

JCP: Proximal Gradient Method

Let $h(\Omega)$ be a lower semi-continuous convex function of Ω, which is not identically equal to $+\infty$. Then proximal point algorithm [Rockafellar [1976]] generates a sequence of solutions $\{\Omega_k,\ k = 1, 2, 3, ...\}$ to the following optimization problem,

$$\Omega_k = Prox_h(\Omega_{k-1}) = \arg\min_{\Omega \succ\ 0} \left(h(\Omega) \ + \ \frac{1}{2}\|\Omega - \Omega_{k-1}\|_2^2 \right). \tag{0.8}$$

The sequence $\{\Omega_k,\ k = 1, 2, 3, ...\}$ weekly converges to the optimal solution of $\min_{\Omega \succ 0}\ h(\Omega)$ (Rockafellar [1976]). To use the structure of the above optimization algorithm, we use quadratic approximation of $f(\Omega)$, which is justified since f is strictly convex.

JCP: Proximal Gradient Method

Basic Approximation Model

For any $L > 0$, consider the following quadratic approximation model of $f(\Omega)$ at $\Omega^{'}$:

$$Q_L(\Omega, \Omega^{'}) := f(\Omega^{'}) + \ < \Omega - \Omega^{'}, \bigtriangledown f(\Omega^{'}) > + \frac{L}{2}\|\Omega - \Omega^{'}\|_2^2 \qquad (0.9)$$

where <A,B> is the inner product of A and B, and L is a positive constant.

Proximal Gradient Operators

- ℓ_1**-norm:**
  Let $M \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$. The proximal operator of $\|.\|_1$ with constant λ is given by

  $$Prox_{\lambda\|\cdot\|_1}(M) = \text{sign}(M)\max(abs(M) - \lambda, 0) \ , \ \ \lambda > 0.$$

- **Trace norm:**
  Let $M = U\Sigma V^T$ be singular value decomposition of M. Then proximal operator of $\|.\|_*$ with constant τ is given by

  $$Prox_{\tau\|\cdot\|_*}(M) = U\Sigma_\tau V^T,$$

  where Σ_τ is diagonal matrix with $((\Sigma_\tau))_{ii} = max(0, \Sigma_{ii} - \tau)$, and $\tau < \min_{i\leq p}(\Sigma_{ii}) - \epsilon$ for some $\epsilon > 0$.

JCP: Simulation Analysis

Block Type Precision Matrix

Table: Average Relative Error with Standard Error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
JCP	0.0865(0.005)	**0.013(0.003)**	0.0641(0.0024)
Graphical Lasso	0.1891(0.0076)	0.092(0.0044)	**0.0085(0.001)**
SPICE	**0.029(0.006)**	0.0279(0.0036)	0.0669(0.0019)
p=100			
JCP	0.1132(0.0039)	**0.019(0.002)**	0.0623(0.0013)
Graphical Lasso	0.3732(0.0043)	0.2131(0.0028)	**0.0345(0.001)**
SPICE	**0.028(0.003)**	0.0458(0.0048)	0.0729(0.0022)
p=200			
JCP	0.1844(0.0064)	**0.048(0.003)**	**0.048(0.002)**
Graphical Lasso	0.715(0.0171)	0.4275(0.0023)	0.1248(0.0014)
SPICE	**0.07(0.004)**	0.0493(0.0051)	0.0904(0.0013)

JCP: Simulation Analysis

Hub Graph Type Precision Matrix

Table: Average Relative Error with Standard Error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
JCP	0.0795(0.0031)	0.0421(0.002)	**0.012(0.001)**
Graphical Lasso	0.0786(0.0043)	0.049(0.003)	0.016(0.001)
SPICE	**0.0103(0.001)**	**0.01(0.001)**	0.0137(0.0008)
p=100			
JCP	0.137(0.005)	0.0714(0.0031)	0.021(0.0005)
Graphical Lasso	0.177(0.006)	0.1001(0.004)	0.036(0.001)
SPICE	**0.023(0.001)**	**0.008(0.001)**	**0.016(0.001)**
p=200			
JCP	0.229(0.0014)	0.1274(0.0008)	0.0415(0.0003)
Graphical Lasso	0.343(0.003)	0.2121(0.001)	0.075(0.0004)
SPICE	**0.06(0.001)**	**0.034(0.001)**	**0.003(0.0003)**

Summary

- **Broadly Applicable New Method:** JCP covariance matrix estimation allows simultaneous estimation of sparse and well-conditioned inverse covariance matrices. Compared to other methods, it is more flexible in penalizing the over-dispersion in sample eigenvalues.

- **Broadly Applicable Algorithm:** The proposed algorithm is can be used to solve number of problems in covariance matrix estimation, multi-task learning. A limitation of the proposed algorithm is that it is slow for large p due to non-smooth penalty constraints.

Summary

- **Broadly Applicable New Method:** JCP covariance matrix estimation allows simultaneous estimation of sparse and well-conditioned inverse covariance matrices. Compared to other methods, it is more flexible in penalizing the over-dispersion in sample eigenvalues.

- **Broadly Applicable Algorithm:** The proposed algorithm is can be used to solve number of problems in covariance matrix estimation, multi-task learning. A limitation of the proposed algorithm is that it is slow for large p due to non-smooth penalty constraints.

My Research

Publications:

- **Maurya, Ashwini**, A Well-conditioned and Sparse Estimation of Covariance and Inverse Covariance Matrices using a Joint Penalty, *Journal of Machine Learning Research Vol 15-345, 2016*.

- **Maurya, Ashwini**, A Joint Convex Penalty for Inverse Covariance Matrix Estimation, *Computational Statistics and Data Analysis, Volume 74, July 2014, 15-27*.

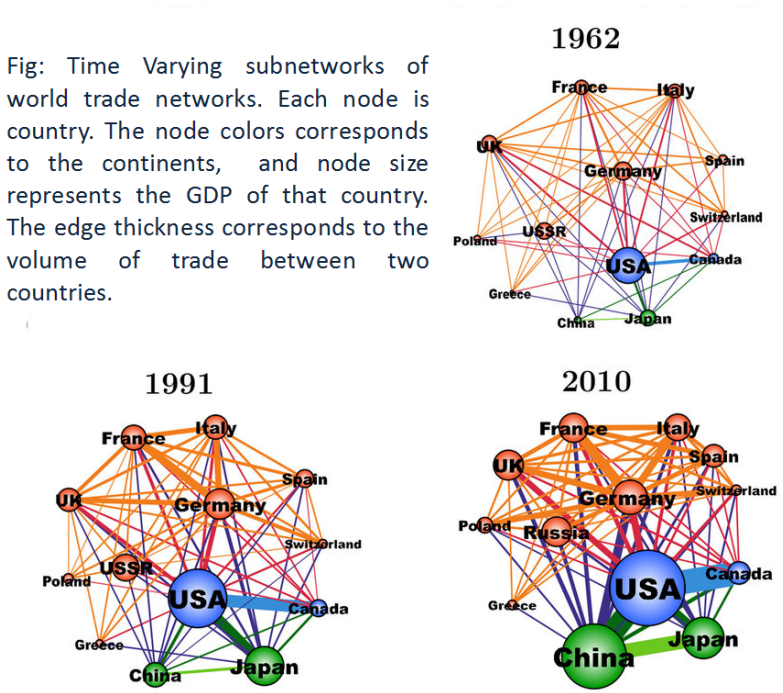
Future Work Directions

Future Work and Extensions

1. Spatio-Temporal Covariance Matrix Estimation:

In time varying (dynamic) setup, the data observed at different time points and locations are not independent.

Figure: Sub-networks of World Trade Network Over Time



Future Work and Extensions

Applications:

- Statistical genetics: Many of the quantitative traits are dynamic in nature. For example the estimation of the dynamic structure of complex traits will advance our knowledge of how the gene networks evolve over time.
- Network Analysis: Estimation of traffic networks, cellular networks, world trade networks.

2. **Prediction of Time Varying Networks**
   - Prediction of how a diseases spreads over time and space.
3. Consistent Estimation of Eigen-vectors in High Dimensional Setting.

Acknowledgment

- **Dr. Hira L. Koul (Michigan State University, USA)**
  I would like to express my deep gratitude towards his valuable and constructive suggestions during the planning and execution of this research work, and for agreeing to be my adviser at MSU.
- **Dr. Adam Rothman (University of Minnesota, USA)**
  For his valuable discussion and suggestions.
- **Dr. Yuehua Cui, (Michigan State University, USA)**
  For motivating discussions on analyzing genetic data, and for agreeing to be on committee.
- **Dr. Mark A. Iwen, (Michigan State University, USA)**
  For teaching the high dimensional class, for discussions on related topics, and for agreeing to be on committee.
- **Dr. Hyokyoung Hong, (Michigan State University, USA)**
  For her support, and for agreeing to be on committee.
- **Dr. Sandipan Roy (University College London, UK)** For helpful discussions, and codes.

Any Questions ?

Thank you for your attention !

For references and other details, I can be reached at akmaurya07@gmail.edu.

Appendix

Types of Cov Matrices

We consider the following five different types of covariance matrices in our simulations.
(i) Hub Graph: Here the rows/columns of Σ_0 are partitioned into J equally-sized disjoint groups: $\{V_1 \cup V_2 \cup, ..., \cup V_J\} = \{1, 2, ..., p\}$, each group is associated with a pivotal row k. Let size $|V_1| = s$. We set $\sigma_{0i,j} = \sigma_{0j,i} = \rho$ for $i \in V_k$ and $\sigma_{0i,j} = \sigma_{0j,i} = 0$ otherwise. In our experiment, $J = [p/s], k = 1, s + 1, 2s + 1, ...,$ and we always take $\rho = 1/(s + 1)$ with J = 20.
(ii) Neighborhood Graph: We first uniformly sample $(y_1, y_2, ..., y_n)$ from a unit square. We then set $\sigma_{0i,j} = \sigma_{0j,i} = \rho$ with probability $(\sqrt{2\pi})^{-1} exp(-4\|y_i - y_j\|^2)$. The remaining entries of Σ_0 are set to be zero. The number of nonzero off-diagonal elements of each row or column is restricted to be smaller than $[1/\rho]$, where ρ is set to be 0.245.
(iii) Toeplitz Matrix: We set $\sigma_{0i,j} = 2$ for $i = j$; $\sigma_{0i,j} = |0.75|^{|i-j|}$, for $|i - j| = 1, 2$; and $\sigma_{0i,j} = 0$, otherwise.
(iv) Block Toeplitz Matrix: In this setting Σ_0 is a block diagonal matrix with varying block size. For $p = 500$, number of blocks is 4 and for $p = 1000$, the number of blocks is 6. Each block of covariance matrix is taken to be Toeplitz type matrix as in the case (iii).

(v) Cov-I type Matrix: In this setting, we first simulate a random sample $(y_1, y_2, ..., y_p)$ from standard normal distribution. Let $x_i = |y_i|^{3/2} * (1 + 1/p^{1+log(1+1/p^2)})$. Next we generate multivariate normal random vectors $\mathbf{Z} = (z_1, z_2, ..., z_{5p})$ with mean vector zero and identity covariance matrix. Let U be eigenvector corresponding to the sample covariance matrix of $\mathbf{Z}$. We take $\Sigma_0 = UDU'$, where $D = \text{diag}(x_1, x_2, ...x_p)$. This is not a sparse setting but the covariance matrix has most of eigenvalues close to zero and hence allows us to compare the performance of various methods in a setting where most of eigenvalues are close to zero and widely spread as compared to structured covariance matrices in the cases **(i)-(iv)**.

PCA Application: Retail Data

- Database $\mathbf{D}$ of $\mathbf{N}$ customers measured over $\mathbf{T}$ time points.
- N can be very big often in millions and T in thousands.
- Goal is to compress $\mathbf{D}$ into data matrix $\mathbf{C}$ of dimension $N \times r$ where $r << T$, without increasing the compression error.
- This can be done by projecting $\mathbf{D}$ onto vectors(s) $a(s)$, such that the variance of $\mathbf{D}a$ is maximized. This is equivalent to solving following problem:

$$\begin{aligned} &\text{maximize} \quad \big(Da\big)^T\big(Da\big) \qquad \text{subject to} \quad a^T a = 1 \\ &\text{Or} \qquad\qquad (D^T D - \lambda I)a \ = 0 \end{aligned} \tag{0.10}$$

- In other words, a is the eigen-vector of covariance matrix $\mathbf{D^T D}$ and λ is the corresponding eigenvalue .

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

C. Stein. Estimation of a covariance matrix. *Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia*, 1975.

C. Stein. Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences.*, 34:1373âĂŞ1403, 1986.

D. Dey and C. Srinivasan. Estimation of a covariance matrix under stein's loss. *Annals of Statistics*, 13(4):1581–1591, 1985.

S. Lin and M. D. Perlman. A monte carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, 6: 411âĂŞ429, 1985.

J.H. Won, J. Lim, S.J. Kim, and B. Rajaratnam. Condition-number regularized covariance estimation. *Journal of the Royal Statistical Society B*, 75, 2012.

U. Alon, Barkai N., Notterman D., Gish K., Ybarra S., Mack D., and Levine A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by

oligonucleotide arrays. *Proceeding of National Academy of Science USA*, 96(12):6745âĂŞ6750, 1999.

P. Bickel and E. Levina. Regulatized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227, 2008a.

P. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(Mar):2577–2604, 2008b.

T. Cai, Z. Ren, and H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 2015.

J. Bien and R. Tibshirani. Sparse estimation of a covariance matrix. *Biometrica*, 98:807–820, 2011.

A. Rothman. Positive definite estimators of large covariance matrices. *Biometrica*, 99:733–740, 2012.

S. Chaudhuri, M. Drton, and T.S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199âĂŞ216, 2007.

L. Xue, Ma S., and Zou Hui. Positive-definite l1-penalized estimation of large covariance matrices. *Journal of American Statistical Association*, 107(500):983–990, 2012.

M. Fazal. Matrix rank minimization with applications., phd thesis. *Elec. Eng. Dept, Stanford University*, 2002.

R.T. Rockafellar. Monotone operators and proximal point algorithm. *SIAM Journal of Control and Optimization*, 14, 1976.