ESTIMATING COVARIANCE STRUCTURE IN HIGH DIMENSIONS

By

Ashwini Maurya

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics and Probability – Doctor of Philosophy

May 3, 2016

ABSTRACT

ESTIMATING COVARIANCE STRUCTURE IN HIGH DIMENSIONS

By

Ashwini Maurya

Many of scientific domains rely on extracting knowledge from high-dimensional data sets to provide insights into complex mechanisms underlying these data. Statistical modeling has become ubiquitous in the analysis of high dimensional data for exploring the large-scale gene regulatory networks in hope of developing better treatments for deadly diseases, in search of better understanding of cognitive systems, and in prediction of volatility in stock market in the hope of averting the potential risk. Statistical analysis in these high-dimensional data sets yields better results only if an estimation procedure exploits hidden structures underlying the data. This thesis develops flexible estimation procedures with provable theoretical guarantees for estimating the unknown covariance structures underlying data generating process. Of particular interest are procedures that can be used on high dimensional data sets where the number of samples n is much smaller than the ambient dimension p. Due to the importance of structure estimation, the methodology is developed for the estimation of both covariance and its inverse in parametric and as well in non-parametric framework.

This thesis is dedicated to my family.
For their endless love, support, and encouragement.

ACKNOWLEDGMENTS

I am really grateful to many people who helped me achieve doctorate in Statistics. It would not have been possible to pursue PhD in United States if it was not for my parents who spend enormous amount of time and effort in educating me from the earliest stages of my life. They supported me in every step of my career and provided a safety net to freely pursue many different possibilities.

At Michigan State University, I am extremely fortunate to be advised by Professor Hira L. Koul, who taught me how to think about the research problems; I have benefited a lot from his clarity of thought and creative intellect. He has always been constant source of motivation and encouraged me to realize my potential. I am grateful to him for providing the best possible academic environment that enabled me to think independently and grow as a scientific researcher. I also thank his family for the amazing hospitality, which in many ways made me at home away from home while at Michigan State.

I have much to thank other members of my thesis committee as well. Dr. Mark Iwen's course on "Compressive sensing and Big Data" and many discussions proved very helpful in my research work. I am thankful to professor Yuehua Cui and Dr. Grace Hong for serving on my thesis committee and for taking time out from their busy schedule to teach me the importance of good research.

I am very grateful to Professor Tathagata Bandyopadhyay at Indian Institute of Management Ahmedabad, for his love, support and encouraging me to pursue advanced degree from United States. I am also very fortunate to know Professor Arnab Laha at Indian Institute of Management Ahmedabad and thank him for his support and encouragement. At Michigan State, I have learned a lot from teaching of Professor Tapabrata Maiti, and thank his family for the unconditional support.

I am thankful to Sue Watson who has been an ever-present source of help, Kim Schmuecker,

and Andy Hufford for their help during many technical issues at Michigan State.

To all my friends, thank you for your understanding and encouragement in my many, many moments of crisis. Your friendship makes my life a wonderful experience. I cant list all the names here, but you are always on my mind.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1
1.1 Covariance Structure Estimation 2
1.1.1 High Dimensional Covariance Matrix Estimation 3
1.2 Inverse Covariance Matrix Estimation 3
1.3 Thesis Overview 4
1.4 Notation 6

I Estimating Covariance Structure **8**

CHAPTER 2 SAMPLE COVARIANCE MATRIX AND ITS LIMITATIONS 9
2.1 Sample Covariance Matrix 9
2.2 Why Sample Covariance Matrix is NOT Suitable in High Dimensions? 10

CHAPTER 3 LOSS FUNCTIONS FOR COVARIANCE MATRIX ESTIMATION . 13
3.1 Likelihood Based Methods 13
3.2 Frobenius Norm Loss Based Methods 14
3.3 Other Loss Function Based Methods 15

CHAPTER 4 LEARNING SPARSE STRUCTURE 17
4.1 Two Broad Class of Covariance Matrices 17
4.2 Lasso Type Penalty 19
4.3 Discussion: 21

CHAPTER 5 ESTIMATING A WELL-CONDITIONED STRUCTURE 23
5.1 Motivation 23
5.2 Well Conditioned Estimation 23
5.3 Variance of Eigenvalues Penalty 25

CHAPTER 6 LEARNING SIMULTANEOUS STRUCTURE WITH JOINT PENALTY 26
6.1 Motivation 26
6.2 JPEN Framework 28
6.3 Theoretical Analysis of JPEN Estimators 29
6.3.1 Results on Consistency 31
6.4 Generalized JPEN Estimators and Optimal Estimation 33
6.4.1 Weighted JPEN Estimator for the Covariance Matrix Estimation . . 33

CHAPTER 7 AN ALGORITHM AND ITS COMPUTATIONAL COMPLEXITY . . 34
7.1 A Very Fast Exact Algorithm . . . 35
7.1.1 Derivation . . . 35
7.1.2 Choice of Regularization Parameters . . . 36
7.1.3 Choice of Weights . . . 36
7.2 Computational Complexity . . . 37

CHAPTER 8 SIMULATIONS . . . 39
8.1 Preliminary . . . 39
8.2 Performance Comparison . . . 42
8.3 Recovery of Eigen-structure and Sparsity . . . 45

II Estimating Inverse Covariance Structure 48

CHAPTER 9 INVERSE COVARIANCE MATRIX AND ITS APPLICATIONS . . . 49
9.1 Motivation . . . 49
9.2 Related Work . . . 50
9.3 Joint Penalty for Precision Matrix Estimation . . . 51
9.4 Some Applications . . . 53
9.4.1 Linear Discriminant Analysis . . . 53
9.4.2 Gaussian Graphical Modeling . . . 53

CHAPTER 10 A JOINT CONVEX PENALTY(JCP) ESTIMATION . . . 55
10.1 Motivation . . . 55
10.2 Joint Convex Penalty Estimation . . . 55
10.2.1 Problem Formulation . . . 56
10.2.2 Proposed Estimator . . . 56

CHAPTER 11 PROXIMAL GRADIENT ALGORITHMS AND ITS CONVERGENCE ANALYSIS . . . 58
11.1 Introduction . . . 58
11.2 Proximal Gradient Method . . . 58
11.3 Basic Approximation Model . . . 59
11.4 Algorithm for optimization . . . 60
11.5 Choosing the Regularization Parameter . . . 61
11.6 Convergence Analysis . . . 61
11.7 Simulation Study . . . 63
11.7.1 Performance Criteria . . . 63
11.7.2 StARS Method of Tuning parameter selection: . . . 64
11.7.3 Simulation Results . . . 64
11.7.3.1 Toeplitz Type Precision Matrix . . . 64
11.7.3.2 Block Type Precision Matrix . . . 66
11.7.3.3 Hub Graph Type Precision Matrix . . . 67
11.7.3.4 Neighborhood Graph Type Precision Matrix . . . 68

11.8 Summary 69
11.9 Discussion 69

CHAPTER 12 SIMULTANEOUS ESTIMATION OF SPARSE AND WELL-CONDITIONED PRECISION MATRIX 71
12.1 Motivation 71
12.2 Joint Penalty Estimation: A Two Step Approach 72
12.3 Weighted JPEN estimator for precision matrix 73
12.4 Theoretical Analysis of JPEN estimators 73

CHAPTER 13 SIMULATIONS AND AN APPLICATION TO REAL DATA ANALYSIS 75
13.1 Simulation Results: Settings 75
13.2 Performance Comparison 76
13.3 Colon Tumor Gene Expression Data Analysis 77

APPENDIX 80

BIBLIOGRAPHY 93

LIST OF TABLES

Table 8.1 Covariance matrix estimation 43

Table 8.2 Covariance matrix estimation 44

Table 11.1 Average KL-Loss with standard error over 20 replications 65

Table 11.2 Average relative error with standard error over 20 replications 65

Table 11.3 Average KL-Loss with standard error over 20 replications 66

Table 11.4 Average relative error with standard error over 20 replications 66

Table 11.5 Average KL-Loss with standard error over 20 replications 67

Table 11.6 Average relative error with standard error over 20 replications 67

Table 11.7 Average KL-Loss with standard error over 20 replications 68

Table 11.8 Average relative error with standard error over 20 replications 68

Table 13.1 Precision matrix estimation 76

Table 13.2 Precision matrix estimation 77

Table 13.3 Averages and standard errors of classification errors over 100 replications in %. 79

LIST OF FIGURES

Figure 2.1 Eigenvalue of sample and population covariance matrices 11

Figure 3.1 A concave function 13

Figure 4.1 Dense and sparse covariance and precision matrices 18

Figure 6.1 Comparison of eigenvalues of covariance matrix estimates 27

Figure 7.1 Timing comparison of JPEN, Glasso, and PDSCE. 38

Figure 8.1 Covariance graph for different type of matrices 41

Figure 8.2 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. 45

Figure 8.3 Eigenvalues plot for $n = 100,\ p = 50$ based on 50 realizations for neighborhood type of covariance matrix 46

Figure 8.4 Eigenvalues plot for $n = 100,\ p = 100$ based on 50 realizations for Cov-I type matrix 47

Figure 9.1 Eigenvalues plot of precision matrix 52

Figure 9.2 Illustration of conditional independence 54

Figure 13.1 Colon tumor gene expression data 78

Figure 13.2 Partial correlation network of colon tumor gene expression data 80

CHAPTER 1

INTRODUCTION

The increasing use of technology with the developments in storage systems has created vast amount of high dimensional data across many scientific disciplines. Examples include the large-scale omics data that enhance our knowledge of human biology, network data that explains how we interact and connect with each other, and the finance data that provides an opportunity to beat the market. The statistical analysis of these data is challenging due to the curse of dimensionality. New statistical methods are needed to model the unknown structure underlying these data sets to leverage our scientific understanding.

The statistical inference in high-dimensional data is possible only if an inference procedure is flexible enough to exploit the hidden structure underlying these data sets. This translates to designing an inference procedure that does well in modeling the structure underlying these data sets. Such an inference procedure often assumes that many of high dimensional structures can be represented with a smaller number of parameters which is the case in many scientific disciplines. Consequently, the the concept of parsimony becomes crucial in high dimensions.

The effectiveness of statistical estimation in high dimensional setting relies on the robustness of the procedure, its efficiency, and scalability. The latter depends upon the availability of scalable algorithmic techniques and its ability to efficiently learn the structure underlying these data sets.

The main goal of this thesis is to develop a flexible and principled statistical methods for uncovering hidden structure underlying the high dimensional, complex data with a focus on scientific discovery. In particular the thesis addresses two main tasks: (i) Estimation of covariance matrices and its inverse, and (ii) Scalable algorithms for computing the covariance structure based on the proposed estimation procedures, in high dimensional setting.

1.1 Covariance Structure Estimation

In many scientific disciplines, a study involves large complex systems whose output often depends on large number of components (variables) and their interactions. As a motivating example, in system biology, the cellular networks often consists of very large number of molecules that interact and exchange information among themselves. Many of the existing techniques depend upon the descriptive analysis of macroscopic properties, which include degree distribution, path lengths and motif profile of these molecular networks or data mining tools to identify clusters. Such an analysis provides limited insight into the complex mechanism of the functional and structural organization of biological structures. An estimate of the robust covariance matrix can explain the biologically important interactions and enhance our scientific understanding of the complex phenomenon.

Covariance structure estimation is of fundamental importance in multivariate data analysis. It is widely used in number of applications including **(**i) Principal component analysis (PCA) [Johnstone and Lu [2004], Zou et al. [2006]], where the goal is to project the data on "best" k-dimensional subspace, and where best means the projected data explains as much of the variation in original data without increasing k; **(ii)** Discriminant analysis [Mardia et al. [1979]]: where the goal is to classify observations into different classes, here estimates of covariance and inverse covariance matrices play an important role as the classifier is often a function of these entities; **(iii)** Regression analysis: If interest focuses on estimation of regression coefficients with correlated (or longitudinal) data, a sandwich estimator of the covariance matrix may be used to provide standard errors for the estimated coefficients that are robust in the sense that they remain consistent under mis-specification of the covariance structure; and **(iv)** Gaussian graphical modeling [Yuan and Lin [2007], Wainwright et al. [2006]Wainright [2009], Yuan [2009],Meinshausen and Bühlmann [2006]]: the relationship structure among nodes can be inferred from inverse covariance matrix (also called precision matrix). A zero entry in the precision matrix implies conditional independence between the

corresponding nodes, given the remaining nodes. In applications where the probability distribution of data is multivariate Gaussian, precision matrix is used to describe the underlying dependece structure among the variables.

1.1.1 High Dimensional Covariance Matrix Estimation

In many of the applications, statistician often encounter data sets where sample size n is much smaller than the ambient dimension p, where the latter can be very large often in thousands, millions or even more. In such situations, the classical statistical estimators tend to have huge bias for estimating their population counterpart and many of the existing asymptotic theories no longer remain valid. In general a p dimensional covariance matrix requires estimation of $p(p+1)/2$ free parameters. For $n < p$, this is an ill-defined problem. In such high dimensional setting, an estimation of covariance matrix is possible by the fact that in many scientific problems, most of the variables are uncorrelated and hence an assumption of sparsity reduces the effective number of parameters to estimate.

1.2 Inverse Covariance Matrix Estimation

Many of scientific studies require estimation of a network structure, in particular the conditional dependence relationships among its nodes (variables). In a large population identifying the true interaction among nodes is generally a very hard problem since number of edges scales quadratically to that of number of nodes n. For a motivating example, consider the estimation of network of neurons. With a significant improvement in technology of measuring neural data, it is now possible to record the spike activity of hundreds of neurons at the same time. A central question in such scientific studies is how do the neurons communicate during a given task? The traditional time and trail-shifting methods suffer from limitation that these do not account for behaviors that encompass multiple distinct structures in the brain. The research [Yatsenko et al. [2015]] shows that the neural spike

pattern can be modeled as a combinnation of sparse precision matrix that accounts for local interaction and a low rank matrix representing the common fluctuations and external inputs. A precision matrix approach is appealing as it offers fexibility in estimating a sparse neural network and also useful for predicting the future states of network of neurons.

Another important application is in Gaussian graphical modeling. In Gaussian graphical models, the conditional independence relationships among the nodes (variables) is equivalently represented as a matrix of partial correlation coefficients or a precision matrix. Thus the estimation of network is equivalent to estimating a precision matrix which is obtained by maximizing the Gaussian likelihood function of observations. Since the Gaussian likelihood is a concave function of precision matrix, its optimization can be solved by a number of fast algorithms.

The existing estimation methods of inverse covariance matrices mainly focus on estimating the underlying sparse structure of the given data, and do not account for minimizing the over dispersion in the sample eigen-spectrum. The proposed method here for the estimation of inverse covariance matrix addresses this phenomenon by penalizing an overdispersion term of the sample eigenvalues. We consider the estimation of precision matrix in both parametric and non-parametric framework. The former is based on Gaussian likelihood whereas the latter is based on Frobenius norm loss function.

1.3 Thesis Overview

The main focus of this thesis is the estimation of large dimensional covariance matrix and its inverse from limited sample observations, establish the theoretical consistency, and provide a very fast algorithm for computing these estimates.

In part I (chapter 2 - chapter 8), we focus on covariance structure estimation.

- Chapter 2 reviews the classical sample covariance matrix estimator, and its limitations in high dimensional setting. We address these limitations and build on these in

subsequent chapters.

- Chapter 3 reviews the various loss functions used in estimation of covariance matrices and its inverse, the related optimization problems, and highlight their advantages and limitations.

- Chapter 4 reviews the sparse structure learning of covariance matrices in high dimensional setting, two broad classes of covariance matrices, and describes their estimation paradigms. We introduce regularized estimation of covariance matrices and discuss their estimation framework with different loss functions from earlier chapter.

- Chapter 5 reviews the concept of well-conditioned estimation and its importance in high dimensional settings. Some of the eigenvalues shrinkage methods, their advantages, and limitations are described. We conclude the chapter by introducing the new penalty (variance of eigenvalues) to reduce the over-dispersion in the sample eigen-structure and the corresponding estimation frameworks based on Frobenius norm loss function.

- In Chapter 6, we propose the Joint Penalty estimation method. We discuss its asymptotic properties and rates of convergence in Frobenius and spectral norm. We also give generalized joint penalty estimators of covariance matrix in high dimensional settings.

- Chapter 7 contains a derivation of a very fast algorithm for computing the proposed joint penalty estimator and compares its computational time with other some existing methods.

- In Chapter 8, we discuss the extensive simulation analysis to compare the performance of the proposed estimator with some other existing methods for various choices of covariance matrices in high dimensional setting. We also analyze the recovery of true sparse and eigen-structure based on joint penalty methods. We conclude the chpater with the classification analysis of sonar data.

In part II (chapter 9- chapter 13), we focus on precision matrix estimation.

- Chapter 9 reviews the existing literature and introduces the proposed joint penalty framework for precision matrix estimation. The chapter concludes with discussion of few applications.

- In Chapter 10, we describe the joint convex penalty (JCP) framework of precision matrix estimation and discuss sparse and well-conditioned estimation under assumption that the data generating process is Gaussian.

- Chapter 11 reviews the proximal gradient algorithm, and its convergence analysis. We give a fast algorithm for computing the JCP estimate of precision matrix. The chapter is concluded with extensive simulation analysis for varying sample sizes and dimensions in high-dimensional setting.

- In Chapter 12, we review simultaneous estimation of sparse and well-conditioned inverse covariance matrices in high dimensional settings. We introduce weighted joint penalty estimators and discuss their rates of convergence in both the Frobenius and spectral norm.

- Chapter 13 reviews the extensive simulation analysis using JPEN method for various choices of structured inverse covariance matrices. We conclude the chapter with an application to colon tumor gene expression data.

Conclusions and future work directions are given in chapter 14.

1.4 Notation

Notation: For a matrix M, M_{ij} denotes its $(i,j)^{th}$ element, $\|M\|_1$ denotes its ℓ_1 norm defined as the sum of absolute values of the entries of matrix M, $\|M\|_F$ denotes the Frobenius

norm of matrix M defined as sum of squared element of M, $\|M\|$ denotes the operator norm (also called spectral norm) defined as largest absolute eigenvalue of M, M^- denotes matrix M where all diagonal elements are set to zero, M^+ denotes matrix M where all off-diagonal elements are set to zero, $\sigma_i(M)$ denotes the i^{th} largest eigenvalue of M, $tr(M)$ denotes its trace, $\|M\|_*$ denotes its trace norm defined as sum of its singular values, and $det(M)$ denotes its determinant.

Part I

Estimating Covariance Structure

CHAPTER 2

SAMPLE COVARIANCE MATRIX AND ITS LIMITATIONS

2.1 Sample Covariance Matrix

Given random vectors $(X_1, X_2, \cdots, X_n)$ from a p-variate probability distribution, the sample covariance matrix is given by:

$$S = [[S_{ij}]], \qquad S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j), \quad i, j = 1, 2, \cdots p. \tag{2.1.1}$$

Sample covariance matrix is a widely applicable estimator of its population counterpart and has low computational complexity. In low dimensional setting where sample size is significantly larger than the number of variables, it possess number of desired properties of a good estimator such as:

- It is theoretically consistent, which means that as the sample size diverges to infinity, it converges almost surely to the population covariance matrix as long as the dimension p is fixed.
- It is an unbiased estimator of its population counterpart.
- It is approximate maximum likelihood estimator.
- Together with the vector of sample means, the sample covariance matrix constitute sufficient statistics for the family of Gaussian distributions.
- 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.

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

2.2 Why Sample Covariance Matrix is NOT Suitable in High Dimensions?

In high dimensional setting where often the dimension exceeds the sample size, typically former is of exponential order of later, many of these properties of sample covariance matrix do not hold. In high dimensions it has following limitations:

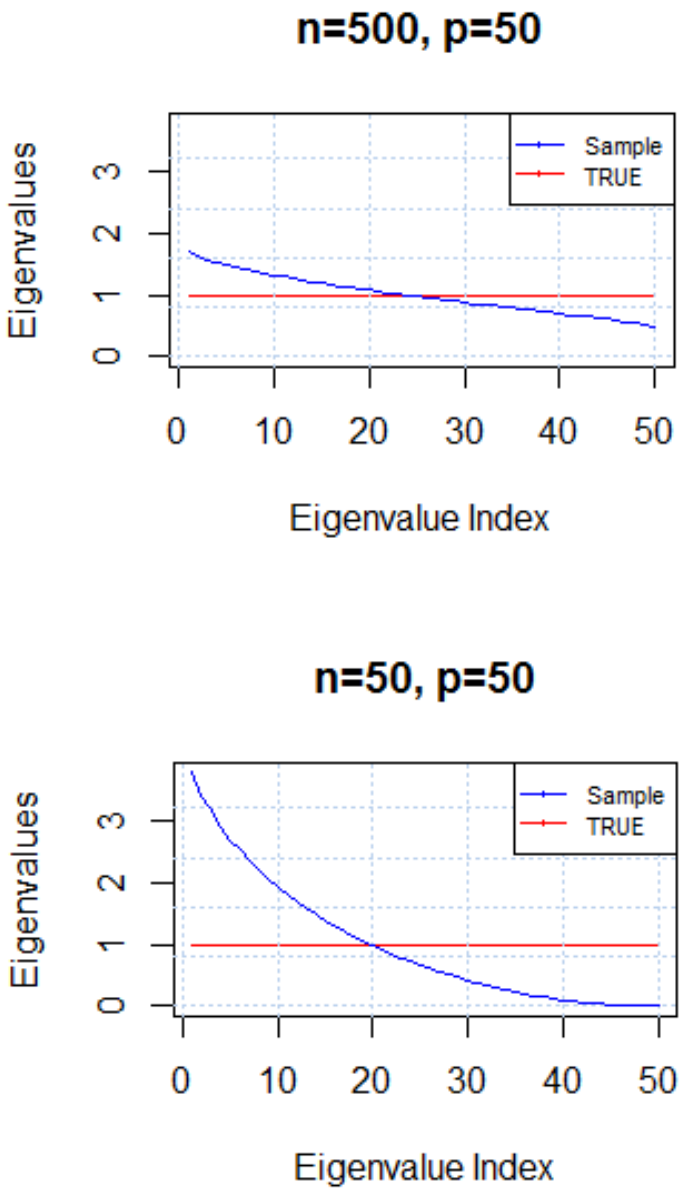
- It is very noisy, which means that the many of its entries have biases.
- For $n < p$, it does not remains positive definite and invertible.
- It has $p - n$ eigenvalues equal to zero which means that total variation in data is contained in first n eigenvalues and therefore highly skewed and biased. In fact the sample eigenvalues are over dispersed in the sense that smaller eigenvalues are biased downward and larger eigenvalues are biased upward of the true eigenvalues.

Figure 2.1 explains the eigen-spectrum over-dispersion phenomenon. For this example, we simulated random vectors from multivariate Gaussian distribtuion with mean zero and identity covariance matrix. We consider two cases:

- case (i) Low dimensional setting: n=500, p=50, and
- case (ii) High dimensional setting n=50, p=50.

The over dispersion in sample eigenvalues are quite apparent in these two settings. The true eigenvalues are all one, whereas the sample eigenvalues follow Marchenko-Pastar law [Marcenko and Pastur [1967]]. A result from [Geman [1980]] shows that for independent and identically (iid) distributed random variables (that have mean zero and identity covariance matrix) with finite fourth moment, as ratio $\frac{p}{n} \to \gamma$, the smallest and largest sample eigenvalues satisfy:

Figure 2.1: Eigenvalue of sample and population covariance matrices



$$l_1 \rightarrow (1+\sqrt{\gamma})^2 \qquad a.s. \qquad \text{and} \tag{2.2.1}$$

$$l_p \rightarrow (1-\sqrt{\gamma})^2 \qquad a.s. \tag{2.2.2}$$

In this example, for case (ii), $\gamma = 1$ by (2.2.1) and (2.2.2), the smallest and largest eigenvalue of sample covariance matrix are 0 and 4 respectively. This shows that in high dimension the sample eigenvalues are over dispersed compared to its population counterpart. Because of these limitations, sample covariance matrix is not a suitable estimator in high dimensional settings.

CHAPTER 3

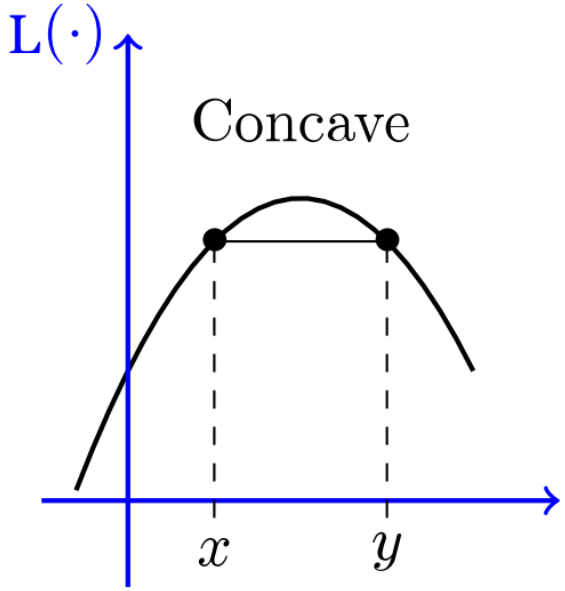
LOSS FUNCTIONS FOR COVARIANCE MATRIX ESTIMATION

Loss functions are key to estimation problems. An estimator optimal with respect to one loss function may not be optimal for other choices of loss functions. The consistency and rate of convergence of the estimators depends upon the choice of loss function. In this chapter we discuss some of the most commonly used loss functions in the context of estimating a high dimensional covariance matrix.

3.1 Likelihood Based Methods

Data likelihood (model based) functions are one of the most widely used loss functions for covariance matrix estimation.

Figure 3.1: A concave function



The likelihood based methods have advantage that often they outperform the non-

parametric counterparts in rate of convergence and asymptotic optimality. In practice, it is reasonable to assume that the data generating likelihood function is smooth. If the likelihood function is strictly concave (Figure 3.1), the unique maximum likelihood estimator exists. In such cases, maximum likelihood estimator can be easily computed using very fast numerical algorithm such as Expectation Maximization algorithm, and the algorithms based on linear or quadratic approximations of likelihood functions. Another advantage of likelihood based estimator is that this can be easily generalized when samples come from a mixture of probability distributions.

Multivariate normal distribution is the most widely used parametric model for covariance matrix estimation. Let $(X_1, X_2, \cdots, X_n)$ follow p-variate normal distribution with zero mean vector and covariance matrix Σ. The likelihood function is given by:

$$L(X_1, X_2, \cdots, X_n; \mu, \Sigma) = \frac{1}{(2\pi)^{np/2}} \frac{1}{|\Sigma|^{p/2}} exp\{-\frac{1}{2} tr(S\Sigma^{-1})\} \tag{3.1.1}$$

This is concave in Σ^{-1}. Therefore a common practice is to maximize the above function with respect to Σ^{-1}. Let $\hat{\Omega}$ be its solution, then an estimate of covariance matrix is given by $\hat{\Omega}^{-1}$.

3.2 Frobenius Norm Loss Based Methods

Frobenius loss function is one of the most popular alternative estimation method to likelihood based methods. It provides a flexible estimation framework and has number of attractive features such as:

- It is convex and easy to solve.
- It is fully non-parametric and does not require any knowledge of functional form of underlying data distribution
- Since the parameter of interest appears in very simple form in the loss function, it is easy to interpret.

- The convex structure facilitates the estimation procedure and its computation can be easily performed with a number of fast algorithms with low computational complexity.

One of most important advantage is that the Frobenius loss function results in exact optimization, unlike in the case of Gaussian distribution, a direct maximization of the function (3.1) with respect to Σ is very difficult problem due to its convex nature. Another disadvantage of likelihood based model is that if data does not meet the stated model assumption (or if model is not choosen carefully), estimators based on likelihood models tend to perform worse than non-parametric estimators. The likelihood function may not always be a concave function, which makes the computation very difficult. In this case, the estimators do not remain optimal anymore.

Let S be the sample covariance matrix. A, un-regularized covariance matrix estimator based on minimization of Froebnius loss function is given by:

$$\hat{\Sigma} = \arg\min_{\Sigma} \|\Sigma - S\|_F^2 \tag{3.2.1}$$

The estimator is $\hat{\Sigma} = S$. As discussed earlier, S may not be suitable estimator in high dimensional setting, and a number of techniques namely regularization and positive well conditioned estimation are needed to improve the sample covariance estimator. For more details on this topic see chapters 4 and 5.

3.3 Other Loss Function Based Methods

The likelihood based methods require a priori knowledge about the underlying data generating stochastic process which is a very strong assumption. Also both the likelihood function based method and Frobenius norm loss function based methods assume the prior knowledge of sample covariance matrix S. In practice sample covariance matrix may not be readily available but some structure of S may be known. In such situations entropy loss function and quadratic loss function are two other commonly used loss functions for covari-

ance matrix estimation [Donoho et al. [2015]].

Entropy loss function: Given a covariance matrix A, not necessarily sample covariance matrix, we seek estimate $\hat{\Sigma}$ which is obtained my minimizing the following entropy loss function:

$$\hat{\Sigma} \quad = \arg\min_{B \succ 0, B=B^T} \quad \left[tr(A^{-1}B) - \log(\det(A^{-1}B)) - p \right] \tag{3.3.1}$$

where A is symmetric and positive definite. The entropy loss function (also known by Kullback-Leibler loss, or Stein's loss function), is a widely used method to measure the discrepancy between two probability distributions.

Quadratic Loss Function An estimator of covariance matrix based on minimization of the quadratic loss function is given by:

$$\hat{\Sigma} = \underset{B \succ 0, B=B^T}{\arg\min} \qquad \left[(tr(A^{-1}B - I)) \right] \tag{3.3.2}$$

The estimators based on entropy and quadratic loss function work well in low dimensional setting when $n < p$.

Remark 3.3.1. Although estimators based on minimizing the Frobenius, entropy and quadratic loss functions have many good properties, to establish the rate of convergence of these estimators, it is necessary to assume some parametric structure on the underlying data generating process. One of the most common assumption is that the data generating process is sub-Gaussian. A continuous random vector is sub-Gaussian if its tails are similar those of Gaussian random vector. See §6.3 for more details.

CHAPTER 4

LEARNING SPARSE STRUCTURE

As discussed in chapter 2, sample covariance matrix is not a suitable estimator in high dimensional setting as this fails to exploit the sparse structure of the true covariance matrix. In this chapter, we discuss improved estimation of sparse covariance matrices based on regularized loss function optimization.

4.1 Two Broad Class of Covariance Matrices

In real data analysis problems, knowledge of true covariance structure is often unknown. However in these situations a suitable assumption on the covariance matrix structure facilitates the estimation procedure and reduces the computational complexity. As a motivating example, given the weather temperature of locations across some geography, we expect the temperature of nearby locations to be similar than the far away locations. Toeplitz type covariance matrix would be suitable for modeling the tempreture variations for that geography.

The most commonly used covariance matrix structures across many scientific disciplines can be classified into two broad class:

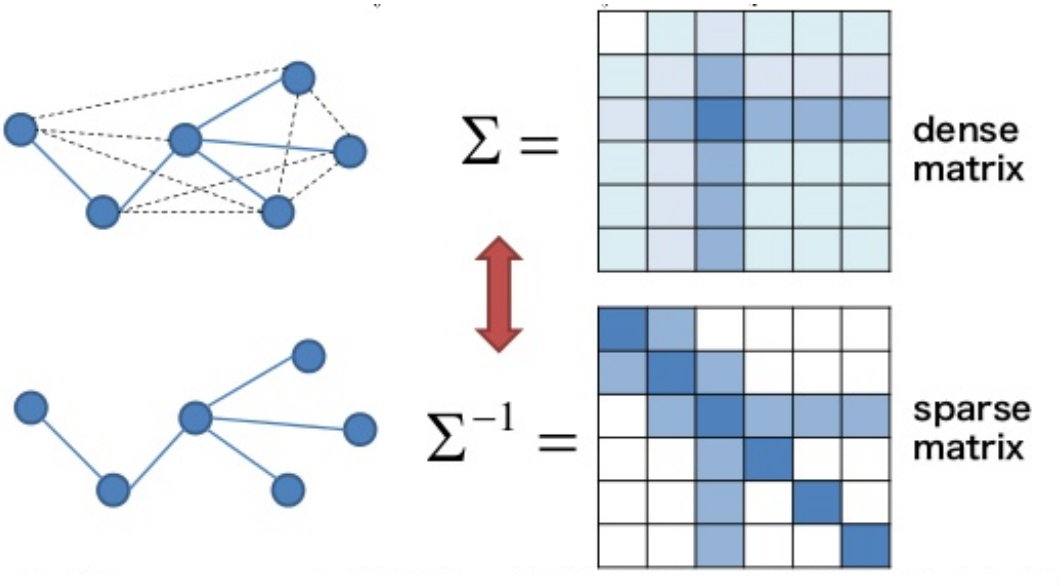
1. Natural ordering among variables: This class includes the covariance matrices where the variables far apart are weakly correlated. One of the example is in time series analysis, where the observations are typically auto correlated in time. In these applications the researcher often assumes that the true underlying covariance matrix has Toeplitz type of structure. Such an assumption greatly reduces the effective number of parameters to be estimated in the matrix.

2. No natural ordering among variables: This class includes the covariance ma-

trices where there is no natural ordering. Examples include the analysis of gene expression data where prior knowledge of any canonical ordering is not available and searching over all permutations of variables is quite infeasible.

In high dimensional setting typically $n < p$, and the estimation of $p(p+1)/2$ free parameters of the covariance matrix based on n observations is ill defined problem. The concept of sparse structural assumption, where one often believe that only few of the entries of true covariance matrix are non-zero, greatly reduces the effective number of parameters to be estimated and hence improves the overall estimation. Figure 4.1 shows the difference between parsimonious (sparse) and non-parsimonious (dense) covariance matrices.

Figure 4.1: Dense and sparse covariance and precision matrices



There is an extensive literature on the estimation of sparse covariance matrix [Bickel and Levina [2008a], Bickel and Levina [2008b], Bien and Tibshirani [2011], Rothman [2012], Xue et al. [2012],Dahl et al. [2008]. Among earlier developments, Dempster [1972] introduced the concept of covariance selection in the context of precision matrix estimation. His approach is based on entrywise sparse estimation of the precision matrix. He shows that the resulting estimator corresponds to maximum entropy model (maximum entropy model is maximum

smooth model among a class of given models). One can follow the similar procedure for covariance matrix estimation in high dimensonal setup by setting certain elements of S to equal zero and continue doing so untill there is no substantial improvement in model fitting. In such a setting it may not be possible to derive an exact test of significance, however number of approximate methods such as change in 2 log Lik value can be used as stopping criteria. The main limitation of such a procedure is that the resulting matrix may not remain positive definite.

The methods for the estimation of high dimensional sparse covariance matrix tend to impose certain structures as suitable on a given class of covariance matrices. For the class of covariance matrices where the variables are assumed to have natural ordering, estimators based on banding or tapering seem to be a natural choice. Bickel and Levina [2008b] proposed regularized estimation of covariance matrices based on banding where the corresponding estimator is obtained by selecting at most k non zero elements in each row. A choice of k is made based on re-sampling and cross validation. Although their estimator has natural interpretation, but need not be positive definite. To overcome this, they propose a tapering estimator of covariance matrices, which uses the Shur matrix multiplication. This is based on the fact that Shur matrix multiplication of two positive definite matrices is also positive definite. For more discussion on this see Bickel and Levina [2008b], Cai et al. [2010], and Karoui [2008].

4.2 Lasso Type Penalty

In situations where there is no natural ordering among variables, banding and tapering based estimators fail to recover the sparse structure of underlying true covariance matrix. In these situations the ℓ_1 regularized covariance matrix estimators are generally permutation invariant [Rothman et al. [2008]] and better alternative than their banding and tapering counterparts. The ℓ_1 based regularized covariance matrix estimation is motivated by lasso

in regression [Tibshiran [1996]], where the main idea is to shrink the smaller entries of covariance matrix to zero, while preserving the positive definiteness. This procedure is also known as covariance graph estimation of marginally independent variables.

Among the likelihood based methods, Bien and Tibshirani [2011] proposed an estimator of covariance matrix as the solution to following optimization problem:

$$\hat{\Sigma} = \underset{\Sigma \succ 0}{\arg\min} \left[log(det(\Sigma)) + tr(S\Sigma^{-1}) + \lambda * \|H * \Sigma\|_1 \right] \tag{4.2.1}$$

where λ is some positive tuning parameter, H is a matrix of non-negative weights. Both λ and H together control the level of sparsity in the estimated covariance matrix. As the above optimization problem is concave in Σ, they derive a solution of (4.2.1) by iteratively solving its convex approximation using "Majorization-Minimization" approach. However, such a procedure is computationaly intense and need not be globally optimal. Among other related works, Chaudhuri et al. [2007] consider the problem of estimating a covariance matrix given a pre-specified zero pattern, Khare and Rajaratnam, in an unpublished 2009 technical report available at http://statistics.stanford.edu/c̃kirby/techreports/GEN/ 2009/ 2009-01.pdf, formulate a prior for Bayesian inference given a covariance graph structure, and Butte et al. [2000] introduce the related notion of relevance network, where genes with partial correlation exceeding given a threshold are connected.

Among Frobenius norm based estimation of high dimensional covariance matrix, Rothman [2012] proposed a correlation matrix estimator as the solution to the following optimization problem:

$$\hat{\Gamma} = \underset{\Gamma = \Gamma^T}{\arg\min} \left[\|\Gamma - R\|_F^2 + \lambda * \|\Gamma^-\|_1 - \gamma log(det(\Gamma)) \right] \tag{4.2.2}$$

where R is sample correlation matrix, γ is some constant that ensures the positive definiteness of $\hat{\Gamma}$. The log-determinant barrier is a valid technique to achieve positive definiteness but it is still unclear whether the iterative procedure proposed in Rothman [2012] actually finds the right solution to the corresponding optimization problem. In another interesting paper,

the authors in Xue et al. [2012] proposed an estimator of covariance matrix as a minimizer of penalized Frobenius norm loss function over set of positive definite matrices. Their estimator is positive definite but whether it overcomes the over-dispersion of the sample eigenvalues, is hard to justify.

In another interesting line of work Lam and Fan [2009] proposed a regularized covariance matrix estimator using a non-convex penalty. They propose their estimator for a class of hard-thresholding and SCAD (Smoothly Clipped Absolute Deviation) penalty. The hard-thresholding penalty is given by: $p_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2 \mathbf{1}_{\{\theta < \lambda\}}$, whereas the SCAD penaly is given by: $p_\lambda(\theta) = \lambda \mathbf{1}_{\{\lambda \leq \theta\}} + (a\lambda - \theta)_+ \mathbf{1}_{\{\theta > \lambda\}}/(a-1)$, for some $a > 2$. The main idea behind the non-convex penalty is to reduce the bias when the value of parameter has relatively larger magnitude. For example, the SCAD penalty remains constant when θ is large, whereas the ℓ_1 penalty grows linearly with θ. The main limitations of non-convex penalty is that the proposed algorithms uses iterative procedure based on local convex approximations hence computatinaly intensive. Also it is hard to say if the proposed algorithm converges to the global minima/maxima.

The proposed Joint PENalty (JPEN) method in this thesis uses Frobenius norm loss function and joint penalty of ℓ_1 and variance of eigenvalue of underlying covariance matrix. The choice of squared loss function allows sparse estimation of covariance matrix (rather the sparse precision matrix), and results in very fast algorithm. We introduce variance of eigenvalues penalty to ensure that the estimated covariance matrix is positive definite. For more details on this, see the chapter 6.

4.3 Discussion:

Assumption of sparsity involves a tradeoff between benefit and cost. In particular in high dimensional data analysis, when entries of covariance matrices are set to zero, the noise due to the error of estimation is generally reduced. On the other hand, errors of misspecification are

introduced. Hence the decision to fit a sparse model comes at trade-off between overfitting and model specification. As noted by Dempster [1972], once the parametric model is adopted, choice of level of sparsity is often settled down, specially when the optimal estimates can easily be computed. However, such optimality provides no gaurantee against the cost of introducing unecessary parameters.

CHAPTER 5

ESTIMATING A WELL-CONDITIONED STRUCTURE

5.1 Motivation

In high dimensional data applications, where an inverse of covariance matrix is used, sample covariance matrix can not be used as generally this is not invertible. By a well-conditioned covariance matrix, we mean that its condition number (ratio of maximum and minimum eigenvalues) is bounded above by a positively finite constant (here the constant is not too large). As pointed out by Ledoit and Wolf [2004], a well-conditioned estimator reduces the estimation error and is a desired propoerty in high dimensional settings. In this chapter, we discuss some of the existing literature on well-conditioned estimation, and introduce variance of eigenvalues penaltly as an effective method for improved eigen-structure estimation.

5.2 Well Conditioned Estimation

The problem of well-conditioned covariance matrix estimation is a long studied subject Stein [1975, 1986], Ledoit and Wolf [2004, 2014], Sheena and Gupta [2003], Won et al. [2012]. It has received considerable attention in high dimensional analysis due to the importance of such estimators in many high dimensional data applications. Among the earlier developments to solve this problem, Stein [1975] proposed his famous class of rotation invariant shrinkage estimators. Here the main idea was to keep the same eigenvectors as that of the sample covariance matrix but shrink the eigenvalues towards the center, in order to reduce the eigenvalues dispersion. Let $S := UDU^T$ be eigen-decomposition of the sample covariance

matrix. Stein's estimator is given by :

$$\hat{\Sigma} = UD^{new}U^T \quad \text{where} \quad D^{new} = diag(d_1^{new}, d_2^{new}, \cdots, d_p^{new}) \tag{5.2.1}$$

with

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

, where $(d_{11}, d_{22}, \cdots, d_{pp})$ is the diagonal of D. This class of estimators is further studied by Haff [1980], Lin and Perlman [1985], Dey and Srinivasan [1985], Ledoit and Wolf [2004, 2014]. Although Stein estimator is considered to be "Gold Standard" [Rajaratnam et al. [2014]], it has a number of limitations including, (i) it is not necessarily positive definite, (ii) assumes normality, and (iii) suitable only for low dimension data analysis when sample size exceeds the dimension. Among the earlier work of eigenvalues shrinkage estimation in high dimensional setting, Ledoit and Wolf [2004] proposed an estimator that shrinks the sample covariance matrix towards identity. Their estimator is given by:

$$\rho_1 \mathbf{S} + \rho_2 I, \qquad \text{where} \quad \rho_1, \rho_2 \quad \text{are estimated from data.} \tag{5.2.2}$$

For ρ large enough, the estimator given by (5.2.2) is well-conditioned but need not be sparse as sample covariance matrix is generally not sparse. In another interesting work, Won et al. [2012] consider maximum likelhihood estimation with a condition number constraint. They solve the following optimization problem:

$$\text{Maximize} \qquad L(S, \Sigma) \qquad \text{subject to } \sigma_{max}/\sigma_{min} \leq \kappa_{max}, \tag{5.2.3}$$

where $L(S, .)$ is likelihood function of multivariate Gaussian distribution given in (3.1.1), and κ_{max} is some positive constant. The estimate Σ of (5.2.3) is invertible if κ_{max} is finite, and well conditioned if κ_{max} is moderate. They consider a value of $\kappa_{max} < 10^3$ to be moderate but its values also depends upon the eigenvalues dispersion of true population covariance matrix.

5.3 Variance of Eigenvalues Penalty

The estimators proposed by Stein [1975], Dey and Srinivasan [1985], Ledoit and Wolf [2004], and Won et al. [2012] are well-conditioned and have been used in a number of applications. However these estimators are not sparse, in addition they have the following limitations:

- The rotation invariant estimators do not change the eigen-vectors and hence they remain inconsistent [Johnstone and Lu [2004]].
- These estimators tend to overestimate the number of non-zeros of the underlying true covariance matrix and its inverse.
- The estimator given in Ledoit and Wolf [2004] results in linear shrinkage of eigenvalues towards those of identity matrix, which may not be optimal criteria, as eigenvalues far from center tend to be heavily biased as compared to the eigenvalues in the center.

The choice of variance of eigenvalues has advantage as it allows more shrinkage of the extreme eigenvalues than the ones in center and therefore non-linearly reduces the bias, and the quadratic term leads to very fast and exact algorithm. See chapter 6 for more detailed analysis of the proposed method.

CHAPTER 6

LEARNING SIMULTANEOUS STRUCTURE WITH JOINT PENALTY

6.1 Motivation

From the discussion in chapter 4, it is understood that learning a sparse structure can be achieved by ℓ_1 regularization. From chapter 5, it is learned that a well-conditioned structure can be achieved by suitably shrinking the sample eigenvalues towards its center. However either of these regularization do not provide a simultaneous treatment of sparse and well-conditioned estimation. For example, consider the estimation of covariance matrix by minimizing the ℓ_1 regularized Frobenius norm loss function:

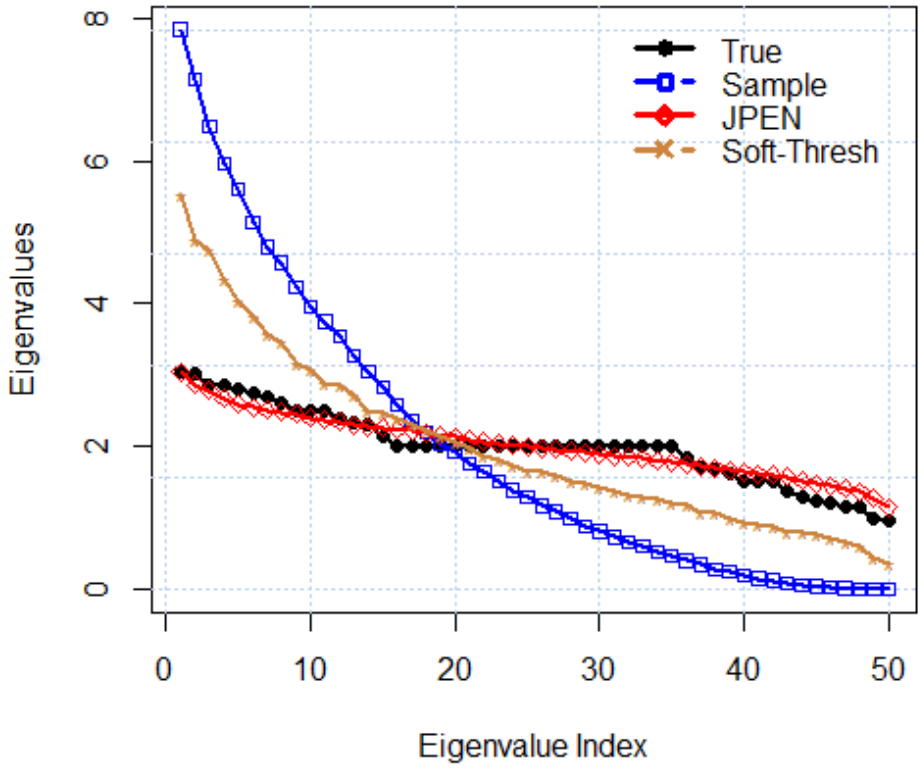
$$\hat{\Sigma}_\lambda = \underset{\Sigma=\Sigma^T,\ tr(\Sigma)=tr(S)}{\arg\min} \Big[||\Sigma - S||_2^2 + \lambda \|\Sigma^-\|_1\Big], \tag{6.1.1}$$

where λ is some positive constant. Note that by penalty function $\|\Sigma^-\|_1$, we only penalize off-diagonal elements of Σ. By the constraint, $tr(\Sigma) = tr(S)$, we ensure that total variation in the estimated covariance matrix is the same as that in the sample covariance matrix. The solution to (6.1.1) is the standard soft-thresholding estimator and it is given by (see chapter 7 for derivation of this estimator):

$$\begin{aligned} \hat{\Sigma}_{ii} &= s_{ii} \\ \hat{\Sigma}_{ij} &= \text{sign}(s_{ij}) \max\Big(|s_{ij}| - \frac{\lambda}{2}, 0\Big), \qquad i \neq j. \end{aligned} \tag{6.1.2}$$

It is clear from this expression that a sufficiently large value of λ will result in sparse covariance matrix estimate. However the estimator $\hat{\Sigma}$ of (6.1.1) is not necessarily positive definite [for more details here see Maurya [2016], Xue et al. [2012]]. Moreover it is hard to say whether it overcomes the over-dispersion in the sample eigenvalues. Figure 6.1 illustrates this phenomenon for a neighbourhood type covariance matrix. Here we simulated random vectors from multivariate normal distribution with sample size $n = 50$ and dimension $p = 50$.

Figure 6.1: Comparison of eigenvalues of covariance matrix estimates



As is evident from Figure 6.1, eigenvalues of sample covariance matrix are over-dispersed as most of them are either too large or close to zero. Eigenvalues of the proposed Joint Penalty (JPEN) estimator are well aligned with those of the true covariance matrix. See chapter 8 for detailed discussion. The soft-thresholding estimator (6.1.2) is sparse but fails to recover the eigenstructure of the true covariance matrix.

To overcome the over dispersion and achieve a well-conditioned estimator, it is natural to regularize the eigenvalues of the sample covariance matrix. Consider the eigenvalues regularized estimator of covariance matrix based on squared loss penalty as the solution to the

following optimization problem:

$$\hat{\Sigma}_\gamma = \underset{\Sigma=\Sigma^T,\ tr(\Sigma)=tr(S)}{\arg\min} \quad \Big[||\Sigma - S||_2^2 + \gamma \sum_{i=1}^{p} \big\{\sigma_i(R) - \bar{\sigma}_R\big\}^2\Big], \tag{6.1.3}$$

where γ is some positive constant. The minimizer $\hat{\Sigma}_\gamma$ of (6.1.3) is given by,

$$\hat{\Sigma} = (S + \gamma\ t\ I)/(1+\gamma), \tag{6.1.4}$$

where I is the identity matrix, and $t = \sum_{i=1}^{p} S_{ii}/p$. To see the advantage of eigenvalue shrinkage penalty, note that after some algebra, for any $\gamma > 0$,

$$\sigma_{min}(\hat{\Sigma}) \quad = \quad \sigma_{min}(S + \gamma\ t\ I)/(1+\gamma) \geq \frac{\gamma\ t}{1+\gamma} > 0.$$

This means that the variance of eigenvalues penalty improves S to a positive definite estimator $\hat{\Sigma}$. However the estimator (6.1.3) is well-conditioned but need not be sparse. Sparsity can be achieved by imposing ℓ_1 penalty on the entries of covariance matrix. In the next section we describe the joint penalty estimation and discuss its advantage as an improved covariance matrix estimator.

6.2 JPEN Framework

We consider the joint penalty estimator of covariance matrix as the solution to the following optimization problem.

$$\hat{\Sigma}_{\lambda,\gamma} = \underset{\Sigma=\Sigma^T,\ tr(\Sigma)=tr(S)}{\arg\min} \quad \Big[||\Sigma - S||_2^2 + \lambda\|\Sigma^-\|_1 + \gamma \sum_{i=1}^{p} \big\{\sigma_i(\Sigma) - \bar{\sigma}_\Sigma\big\}^2\Big], \tag{6.2.1}$$

where λ and γ are some positive constants. From here onwards we suppress the dependence of $\hat{\Sigma}$ on λ, γ and denote $\hat{\Sigma}_{\lambda,\gamma}$ by $\hat{\Sigma}$.

Simulations have shown that, in general the minimizer of (6.2.1) is not positive definite for all values of $\lambda > 0$ and $\gamma > 0$. Therefore we consider the optimization of (6.2.1) for restricted set of (λ, γ) to ensure the resulting estimator is sparse and well-conditioned

simultaneously. In what follows, we first consider correlation matrix estimation, and later generalize the method for covariance matrix estimation.

The proposed JPEN covariance matrix estimator is obtained by optimizing the following objective function in R over specific region of values of (λ, γ) which depends on the sample correlation matrix K, and λ, γ. Here the condition $tr(\Sigma) = tr(S)$ reduces to $tr(R) = p$, and therefore $t = 1$.

$$\hat{R}_K = \underset{R=R^T, tr(R)=p | (\lambda,\gamma) \in \hat{\mathscr{S}}_1^K}{\arg\min} \quad \Big[||R-K||_F^2 + \lambda \|R^-\|_1 + \gamma \sum_{i=1}^{p} \Big\{ \sigma_i(R) - \bar{\sigma}_R \Big\}^2 \Big], \qquad (6.2.2)$$

where

$$\hat{\mathscr{S}}_1^K \quad = \Big\{ (\lambda,\gamma) : \lambda, \gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\tfrac{log\ p}{n}}, \forall \epsilon > 0, \sigma_{min}\{(K+\gamma I) - \tfrac{\lambda}{2} * sign(K+\gamma I)\} > \epsilon \Big\},$$

and $\bar{\sigma}_R$ is the mean of the eigenvalues of R. In particular if K is diagonal matrix, the set $\hat{\mathscr{S}}_1^K$ is given by,

$$\hat{\mathscr{S}}_1^K = \Big\{ (\lambda,\gamma) : \lambda, \gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\tfrac{log\ p}{n}}, \forall \epsilon > 0, \lambda < 2(\gamma - \epsilon) \Big\}.$$

The minimization in (6.2.2) over R is for fixed $(\lambda, \gamma) \in \hat{\mathscr{S}}_1^K$. Furthermore Lemmas 6.2.1 and 6.2.2, respectively show that the objective function (6.2.2) is convex and estimator given in (6.2.3) is positive definite.

The proposed estimator of covariance matrix (based on regularized correlation matrix estimator $\hat{R}_K$) is given by:

$$\hat{\Sigma}_K = (S^+)^{1/2} \hat{R}_K (S^+)^{1/2}, \qquad (6.2.3)$$

where S^+ is the diagonal matrix of the diagonal elements of S.

6.3 Theoretical Analysis of JPEN Estimators

Although we do not make any assumption of data generating process for estimation, to derive rates of convergence we make assumption that the underlying data generating

process is sub-Gaussian. In this section, we give rates of convergence of the proposed JPEN estimator in high dimensional setting where the sample size and dimension both diverge to infinity.

Def: A random vector X is said to have sub-Gaussian distribution if for each $t \geq 0$ and $y \in \mathbb{R}^p$ with $\|y\|_2 = 1$, there exist $0 < \tau < \infty$ such that

$$\mathbb{P}\{|y^T(X - \mathbb{E}(X))| > t\} \leq e^{-t^2/2\tau} \tag{6.3.1}$$

The JPEN estimators exists for any (n,p) satisfying $2 \leq n < p < \infty$. For theoretical consistency in operator norm we require $s \log\ p = o(n)$ and for Frobenius norm we require $(p+s)\log\ p = o(n)$ where s is the upper bound on the number of non-zero off-diagonal entries in true covariance matrix. For more details, see the remark after Theorem 6.3.1.

We make the following additional assumptions about the true covariance matrix Σ_0.

A0. Let $X := (X_1, X_2, \cdots, X_p)$ be a mean zero vector with covariance matrix Σ_0 such that each $X_i/\sqrt{\Sigma_{0ii}}$ has sub-Gaussian distribution with parameter τ as defined in (6.3.1).

A1. With $E = \{(i,j) : \Sigma_{0ij} \neq 0, i \neq j\}$, the $|E| \leq s$ for some positive integer s.

A2. There exists a finite positive real number $\bar{k} > 0$ such that $1/\bar{k} \leq \sigma_{min}(\Sigma_0) \leq \sigma_{max}(\Sigma_0) \leq \bar{k}$.

Assumption A2 guarantees that the true covariance matrix Σ_0 is well-conditioned (i.e. all the eigenvalues are finite and positive). Assumption A1 is more of a definition which says that the number of non-zero off diagonal elements are bounded by some positive integer. The following Lemmas 6.3.1 and 6.3.2, respectively, show that the optimization problem in (6.2.2) is convex and yields a positive definite solution.

Lemma 6.3.1. *The optimization problem in (6.2.2) is convex.*

Lemma 6.3.2. *The estimator given by (6.2.2) is positive definite for any $2 \leq n < \infty$ and $1 \leq p < \infty$.*

6.3.1 Results on Consistency

Theorem 6.3.1. *Let* $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^K$ *and* $\hat{\Sigma}_K$ *be as defined in (6.2.2). Under Assumptions A0, A1, A2,*

$$\|\hat{R}_K - R_0\|_F = O_P\Big(\sqrt{\frac{s\ log\ p}{n}}\Big) \quad and \quad \|\hat{\Sigma}_K - \Sigma_0\| = O_P\Big(\sqrt{\frac{(s+1)log\ p}{n}}\Big), \tag{6.3.2}$$

where R_0 *is true correlation matrix.*

Remark 6.3.1. The JPEN estimator $\hat{\Sigma}_K$ is mini-max optimal under the operator norm. In (Cai et al. [2015]), the authors obtain the mini-max rate of convergence in the operator norm of their covariance matrix estimator for the particular construction of parameter space $\mathscr{H}_0(c_{n,p}) := \Big\{\Sigma : max_{1\le i\le p}\sum_{i=1}^{p} I\{\sigma_{ij}\neq 0\} \le c_{n,p}\Big\}$. They show that this rate in operator norm is $c_{n,p}\sqrt{log\ p/n}$ which is same as that of $\hat{\Sigma}_K$ for $1 \le c_{n,p} = \sqrt{s}$.

Remark 6.3.2. Bickel and Levina [2008b] proved that under the assumption of $\sum_{j=1}|\sigma_{ij}|^q \le c_0(p)$ for some $0 \le q \le 1$, the hard thresholding estimator of the sample covariance matrix for tuning parameter $\lambda \asymp \sqrt{(log\ p)/n}$ is consistent in operator norm at a rate no worse than $O_P\Big(c_0(p)\sqrt{p}(\frac{log\ p}{n})^{(1-q)/2}\Big)$ where $c_0(p)$ is the upper bound on the number of non-zero elements in each row. Here the truly sparse case corresponds to $q = 0$. The rate of convergence of $\hat{\Sigma}_K$ is same as that of Bickel and Levina [2008b] except in the following cases:

Case (i) The covariance matrix has all off diagonal elements zero except last row which has $\sqrt{p}$ non-zero elements. Then $c_0(p) = \sqrt{p}$ and $\sqrt{s} = \sqrt{2\ \sqrt{p}-1}$. The operator norm rate of convergence for JPEN estimator is $O_P\Big(\sqrt{\sqrt{p}\ (log\ p)/n}\Big)$ where as rate of Bickel and Levina's estimator is $O_P\Big(\sqrt{p\ (log\ p)/n}\Big)$.

Case (ii) When the true covariance matrix is tri-diagonal, we have $c_0(p) = 2$ and $s = 2p-2$, the JPEN estimator has operator norm rate of $\sqrt{p\ log\ p/n}$ whereas that of Bickel and Levina's estimator is $\sqrt{log\ p/n}$.

For the case $\sqrt{s} \asymp c_0(p)$ and JPEN estimator has the same rate of convergence as that of Bickel and Levina's estimator.

Remark 6.3.3. The operator norm rate of convergence is much faster than Frobenius norm. This is due to the fact that Frobenius norm convergence is in terms of all eigenvalues of the covariance matrix whereas the operator norm convergence is in terms of the largest eigenvalue.

Remark 6.3.4. Our proposed estimator is applicable to estimate any non-negative definite covariance matrix.

Note that the estimator $\hat{\Sigma}_K$ is obtained by regularization of sample correlation matrix in (6.2.2). In some applications it is desirable to directly regularize the sample covariance matrix. The JPEN estimator of the covariance matrix based on regularization of sample covariance matrix is obtained by solving the following optimization problem:

$$\hat{\Sigma}_S = \underset{\Sigma=\Sigma^T, tr(\Sigma)=tr(S)|(\lambda,\gamma)\in\hat{\mathscr{S}}_1^S}{\arg\min} \Big[||\Sigma - S||_F^2 + \lambda\|\Sigma^-\|_1 + \gamma\sum_{i=1}^{p}\{\sigma_i(\Sigma) - \bar{\sigma}_\Sigma\}^2\Big], \quad (6.3.3)$$

where

$$\hat{\mathscr{S}}_1^S = \Big\{(\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\tfrac{log\ p}{n}}, \forall\epsilon > 0, \sigma_{min}\{(S+\gamma tI) - \tfrac{\lambda}{2} * sign(S+\gamma tI)\} > \epsilon\}.$$

The minimization in (6.3.3) over Σ is for fixed $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^S$. The estimator $\hat{\Sigma}_S$ is positive definite and well-conditioned. Theorem 6.3.2 gives the rate of convergence of the estimator $\hat{\Sigma}_S$ in Frobenius norm.

Theorem 6.3.2. *Let $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^S$, and let $\hat{\Sigma}_S$ be as defined in (6.3.3). Under Assumptions A0, A1, A2,*

$$\|\hat{\Sigma}_S - \Sigma_0\|_F = O_P\Big(\sqrt{\frac{(s+p)log\ p}{n}}\Big) \quad (6.3.4)$$

As noted in Rothman [2012] the worst part of convergence here comes from estimating the diagonal entries.

6.4 Generalized JPEN Estimators and Optimal Estimation

The estimators in (6.2.1) and (6.3.3) encourage eigenvalue shrinkage by the same weights for all the eigenvalues. However one might want to penalize the eigenvalues with different weights, especially if some prior knowledge is available about the structure of true eigenvalues. To encourage different level of shrinkage towards the center, we provide the more generic estimators and call it weighted JPEN estimators.

6.4.1 Weighted JPEN Estimator for the Covariance Matrix Estimation

A modification of estimator $\hat{R}_K$ is obtained by adding positive weights to the term $(\sigma_i(R) - \bar{\sigma}_R)^2$. This leads to weighted eigenvalues variance penalty with larger weights amounting to greater shrinkage towards the center and vice versa. Note that the choice of the weights allows one to use any prior structure of the eigenvalues (if known) in estimating the covariance matrix. The weighted JPEN correlation matrix estimator $\hat{R}_A$ is given by

$$\hat{R}_A = \underset{R=R^T, tr(R)=p | (\lambda,\gamma) \in \hat{\mathscr{S}}_1^{K,A}}{\arg\min} \left[||R-K||_F^2 + \lambda \|R^-\|_1 + \gamma \sum_{i=1}^{p} a_i \{\sigma_i(R) - \bar{\sigma}_R\}^2 \right], \quad (6.4.1)$$

where

$$\hat{\mathscr{S}}_1^{K,A} = \left\{ (\lambda,\gamma) : \lambda \asymp \gamma \asymp \sqrt{\frac{log\ p}{n}}, \lambda \leq \frac{(2\ \sigma_{min}(K))(1+\gamma\ max(A_{ii})^{-1})}{(1+\gamma\ min(A_{ii}))^{-1} p} + \frac{\gamma\ min(A_{ii})}{p} \right\},$$

and $A = \text{diag}(A_{11}, A_{22}, \cdots A_{pp})$ with $A_{ii} = a_i$. The proposed covariance matrix estimator is $\hat{\Sigma}_{K,A} = (S^+)^{1/2} \hat{R}_A (S^+)^{1/2}$. The optimization problem in (6.4.1) is convex and yields a positive definite estimator for each $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^{K,A}$. A simple excercise shows that the estimator $\hat{\Sigma}_{K,A}$ has the same rate of convergence as $\hat{\Sigma}_S$. How to choose weights a_i in (6.4.1), is discussed in next chapter.

CHAPTER 7

AN ALGORITHM AND ITS COMPUTATIONAL COMPLEXITY

A problem is regarded as inherently difficult if its solution requires significant resources, whatever the algorithm used. Despite recent ambitious developments in solving convex optimization problems, efficient computation and scalability still remain two challenging problems in high dimensions data analysis. The existing methods that solve a convex optimization (here we mean minimization) problems often can be implemented very efficiently in far less time than the concave optimization. Extensive literature exits for convex optimization problems [Bertsekas [2010] Vandenberghe and Boyd [2004], Bach et al. [2011], Beck and Teboulle [2009]]. The main challange in covariance matrix estimation in the Gaussian liklihood framework is that the negative of log likelihood is concave function which makes it a very hard optimization problem. In such situations, one way to facilitate the optimization is to approximate the negative of log likelihood function by some non-concave function and then solve this approximate problem efficiently using existing algorithms. However the solution thus obtained may not be an optimal solution to the original problem.

The existing algorithms of computing the optimal covariance matrix based on Frobenius loss function have computational complexity of $O(p^3)$, where the constant in $O(p^3)$ can be really large, often more than the dimension of the matrix. The main reason behind such high computational complexity is that the methods require optimzation over a set of positive definite cones for the estimator to be positive definite (for more on this topic, see Xue et al. [2012]). The JPEN framework provides an easy solution for positive definite constraints that depends upon choices of the (λ, γ). The compuatonal complexity of JPEN estimator is $O(p^2)$ and thus much faster than the other existing algorithms. The next section describes the derivation of JPEN estimator described in chapter 6.

7.1 A Very Fast Exact Algorithm

7.1.1 Derivation

The optimization problem (6.2.2) can be written as:

$$\hat{R}_K = \underset{R=R^T|(\lambda,\gamma)\in\hat{\mathscr{S}}_1^K}{\arg\min} f(R), \tag{7.1.1}$$

where

$$f(R) = ||R-K||_F^2 + \lambda\|R^-\|_1 + \gamma\sum_{i=1}^{p}\{\sigma_i(R)-\bar{\sigma}(R)\}^2.$$

Note that $\sum_{i=1}^{p}\{\sigma_i(R)-\bar{\sigma}(R)\}^2 = tr(R^2) - 2\ tr(R) + p$, where we have used the constraint $tr(R) = p$. Therefore,

$$\begin{aligned} f(R) &= \|R-K\|_F^2 + \lambda\|R^-\|_1 + \gamma\ tr(R^2) - 2\ \gamma\ tr(R) + p \\ &= tr(R^2(1+\gamma)) - 2tr\{R(K+\gamma I)\} + tr(K^TK) + \lambda\ \|R^-\|_1 + p \\ &= (1+\gamma)\{tr(R^2) - \frac{2}{1+\gamma}tr\{R(K+\gamma I)\} + (1/(1+\gamma))tr(K^TK)\} + \lambda\ \|R^-\|_1 + p \\ &= (1+\gamma)\{\|R-(K+\gamma I)/(1+\gamma)\|_F^2 + \frac{1}{1+\gamma}tr(K^TK)\} + \lambda\ \|R^-\|_1 + p. \end{aligned}$$

The solution of the above optimization problem is soft thresholding estimator and is given by,

$$\hat{R}_K = \frac{1}{1+\gamma}\ \text{sign}(K) * \text{pmax}\{\text{abs}(K+\gamma\ I) - \frac{\lambda}{2}, 0\} \tag{7.1.2}$$

with $(\hat{R}_K)_{ii} = (K_{ii}+\gamma)/(1+\gamma)$, $\text{pmax}(A,b)_{ij} := \max(A_{ij}, b)$ is elementwise max function for each entry of the matrix A. Note that for each $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^K$, $\hat{R}_K$ is positive definite.

7.1.2 Choice of Regularization Parameters

For a given value of γ, we can find the value of λ satisfying

$$\sigma_{min}\{(K+\gamma I) - \frac{\lambda}{2} * sign(K+\gamma I)\} > 0, \tag{7.1.3}$$

which can be simplified to

$$\lambda < \frac{\sigma_{min}(K+\gamma I)}{C_{12}\ \sigma_{max}(\text{sign}(K))}, \quad C_{12} \geq 0.5.$$

Then $(\lambda, \gamma) \in \hat{\mathscr{S}}_1^K$, and the estimator $\hat{R}_K$ is positive definite. Smaller values of C_{12} yield a solution which is more sparse but may not be positive definite. The optimal values of (λ, γ) were obtained following the approach suggested in Bickel and Levina [2008b] by minimizing the $5-$fold cross validation error

$$\frac{1}{5}\sum_{i=1}^{5} \|\hat{\Sigma}_i^v - \Sigma_i^{-v}\|_1,$$

where $\hat{\Sigma}_i^v$ is JPEN estimate of the covariance matrix based on $v = 4n/5$observations, Σ_i^{-v} is the sample covariance matrix using $(n/5)$ observations.

7.1.3 Choice of Weights

For the optimization problem in (6.4.1), we chose the weights as per the following criteria:

Let $\mathscr{E} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_p)$ be the set of smallest to largest diagonal elements of the sample covariance matrix S.

- Let k be the largest integer such that k^{th} elements of $\mathscr{E}$ is less than 1. Let

$$b_i = \begin{cases} \epsilon_i & \text{for } i \leq k \\ 1/\epsilon_i, & \text{for } k < i. \end{cases}$$

- $A = \text{diag}(a_1, a_2, \cdots, a_p)$, where $a_j = \frac{b_j}{\sum_{i=1}^{p} b_i}$.

Such choice of weights allows more shrinkage of extreme sample eigenvalues than the ones in the center of eigen-spectrum.

7.2 Computational Complexity

The JPEN estimator $\hat{\Sigma}_K$ has computational complexity of $O(p^2)$. This is due to the fact that there there are at most $(p^2 + 2p)$ multiplication, and at most p^2 operations for entry-wise maximum computation. The other existing algorithms Graphical Lasso (Friedman et al. [2008]), and PDSCE (Rothman [2012]) have computational complexity of $O(p^3)$, where the constant of complexity is often very large, mainly due to the iterative nature of convergence. Another advantage of JPEN estimator is that it is an exact solution to the underlying optimization problem. To see the computing time performance, we plot the computational timing of our algorithm and some other existing algorithms including Glasso (Friedman et al. [2008]), PDSCE (Rothman [2012]). Note that the exact timing of these algorithm also depends upon the implementation, platform etc. (we did our computations in R on a AMD 2.8GHz processor).

Figure 7.1: Timing comparison of JPEN, Glasso, and PDSCE.

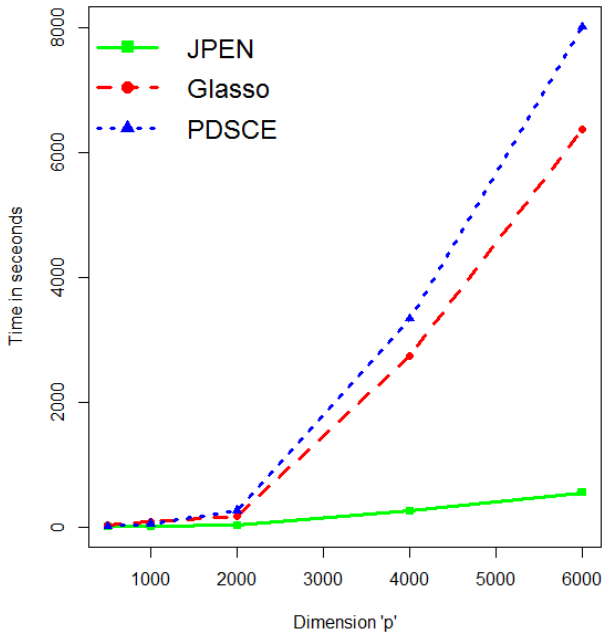


Figure 7.1 illustrates the total computational time taken to estimate the covariance matrix by *Glasso*, $PDSCE$ and $JPEN$ algorithms for different values of p for Toeplitz type of covariance matrix (see chapter 8 for Toeplitz type of covariance matrix). Although the proposed method requires optimization over a grid of values of $(\lambda, \gamma) \in \hat{\mathscr{S}}_1^K$, our algorithm is very fast and easily scalable to large scale data analysis problems.

CHAPTER 8

SIMULATIONS

In this chapter we compare the performance of the proposed JPEN estimator of covariance matrix for various choices of structured covariance matrices. We consider covariance matrices from both class *viz.* (i) when there is a natural ordering among variables, and (ii) when there is no natural ordering among variables. In addition to these, we also include results in a setting when the underlying true covariance matrix is dense and has very high condition number.

8.1 Preliminary

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.

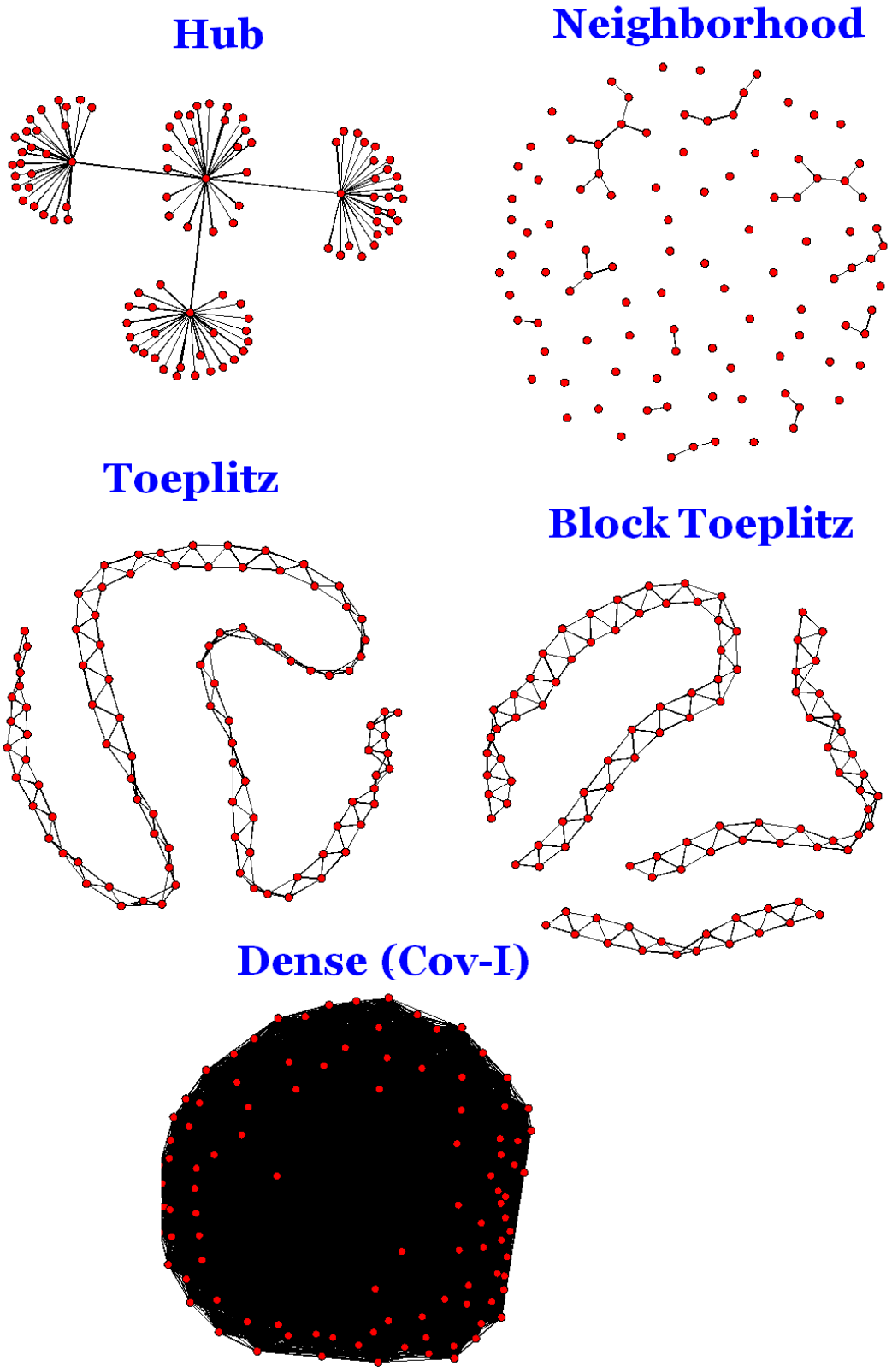
(iii) Toeplitz Matrix: We set $\sigma_{0i,j} = 2$ for $i = j$; $\sigma_{0i,j} = |0.75|^{|i-j|}$, for $|i-j| = 1, 2$; and $\sigma_{0i,j} = 0$, otherwise.

(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 $\mathscr{Z} = (z_1, z_2, ..., z_{5p})$ with mean vector zero and identity covariance matrix. Let U be eigenvector corresponding to the sample covariance matrix of $\mathscr{Z}$. 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)**.

Figure 8.1 shows the graphical covariance structure for these matrices. Here we choose $p = 100$ for better visualization.

Figure 8.1: Covariance graph for different type of matrices



For all these choices of covariance and inverse covariance matrices, we generate random vectors from multivariate normal distributions with varying n and p. We chose $n = 50, 100$ and $p = 500, 1000$. We compare the performance of the proposed covariance matrix estimator $\hat{\Sigma}_K$ with the following estimators.

- **Graphical lasso** [Friedman et al. [2008]]: Graphical lasso estimates a sparse precision matrix. Here we invert the inverse, and include in our analysis. The estimate was computed using 'R' package 'Glasso'. For more details, refer to http://statweb. stanford. edu/ tibs/ glasso/.
- **Bickel and Levina's thresholding estimator (BLThresh)** [Bickel and Levina [2008b]]. The estimator was computed as per the algorithm given in their paper.
- **Rothman's Positive Definite Sparse Covariance Matrix Estimator (PDSCE)** [**Rothman [2012]**]**.** The PDSCE was computed using 'R' package 'PDSCE'. For more details, refer to (http://cran. r-project. org/web/ packages/PDSCE/index.html)
- **Ledoit and Wolf estimator** [Ledoit and Wolf [2004]] Their estimate was computed using code from (http://econ.uzh.ch/faculty/wolf/publications.html#9).
- **The JPEN estimator** was computed using 'R' package 'JPEN'. All the computations were done using R on a AMD 2.8GHz processor.

8.2 Performance Comparison

For each of covariance and precision matrix estimate, we calculate Average Relative Error (ARE) based on 50 iterations using following formula,

$$ARE(\Sigma, \hat{\Sigma}) = |\log(f(S, \hat{\Sigma})) - \log(f(S, \Sigma_0))| / |\log(f(S, \Sigma_0))|, \tag{8.2.1}$$

where $f(S, \cdot)$ is multivariate normal density given the sample covariance matrix S, Σ_0 is the true covariance, $\hat{\Sigma}$ is the estimate of Σ_0 based on one of the methods under consideration.

Other choices of performance criteria are Kullback-Leibler, used by Yuan and Lin [2007] and Bickel and Levina [2008b], precision and recall. The optimal values of tuning parameters were obtained over a grid of values by minimizing $5-$fold cross-validation as explained in §7.2.

Table 8.1: Covariance matrix estimation

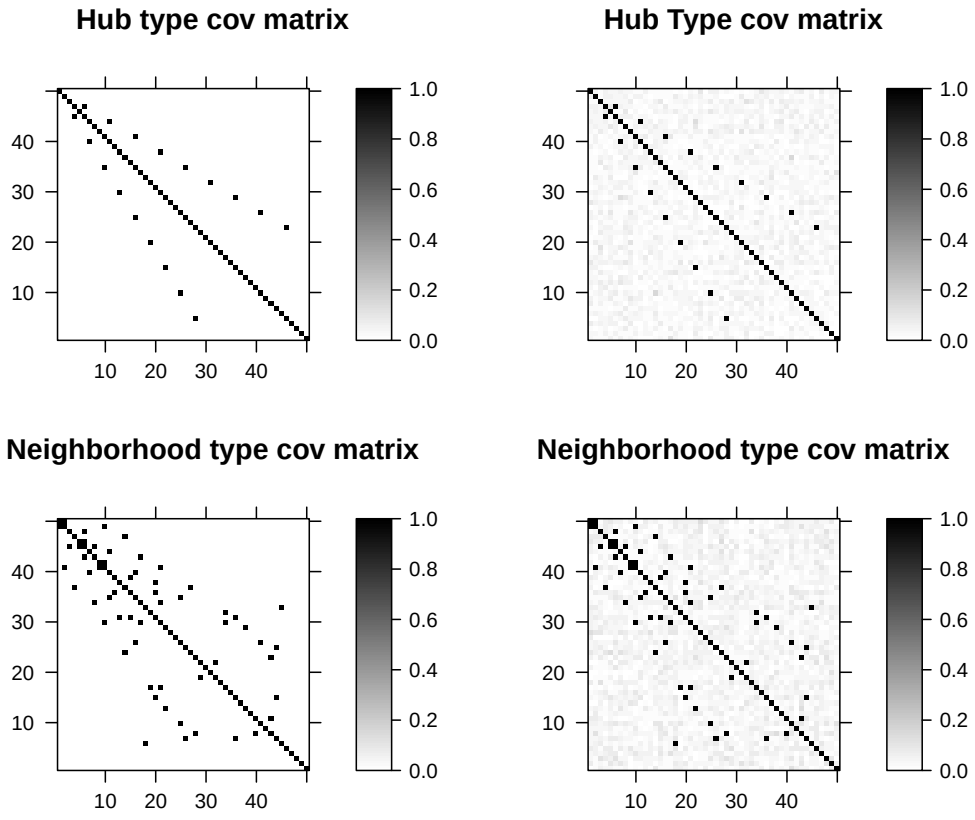
Block type covariance matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Ledoit-Wolf	1.54(0.102)	2.96(0.0903)	4.271(0.0394)	2.18(0.11)
Glasso	0.322(0.0235)	3.618(0.073)	0.227(0.098)	2.601(0.028)
PDSCE	3.622(0.231)	4.968(0.017)	1.806(0.21)	2.15(0.01)
BLThresh	2.747(0.093)	3.131(0.122)	0.887(0.04)	0.95(0.03)
JPEN	2.378(0.138)	3.203(0.144)	1.124(0.088)	2.879(0.011)
Hub type covariance matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Ledoit-Wolf	2.13(0.103)	2.43(0.043)	1.07(0.165)	3.47(0.0477)
Glasso	0.511(0.047)	0.551(0.005)	0.325(0.053)	0.419(0.003)
PDSCE	0.735(0.106)	0.686(0.006)	0.36(0.035)	0.448(0.002)
BLThresh	1.782(0.047)	2.389(0.036)	0.875(0.102)	1.82(0.027)
JPEN	0.732(0.111)	0.688(0.006)	0.356(0.058)	0.38(0.007)
Neighborhood type covariance matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Ledoit-Wolf	1.36(0.054)	2.89(0.028)	1.1(0.0331)	2.32(0.0262)
Glasso	0.608(0.054)	0.63(0.005)	0.428(0.047)	0.419(0.038)
PDSCE	0.373(0.085)	0.468(0.007)	0.11(0.056)	0.175(0.005)
BLThresh	1.526(0.074)	2.902(0.033)	0.870(0.028)	1.7(0.026)
JPEN	0.454(0.0423)	0.501(0.018)	0.086(0.045)	0.169(0.003)

Table 8.2: Covariance matrix estimation

Toeplitz type covariance matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Ledoit-Wolf	1.526(0.074)	2.902(0.033)	1.967(0.041)	2.344(0.028)
Glasso	2.351(0.156)	3.58(0.079)	1.78(0.087)	2.626(0.019)
PDSCE	3.108(0.449)	5.027(0.016)	0.795(0.076)	2.019(0.01)
BLThresh	0.858(0.040)	1.206(0.059)	0.703(0.039)	1.293(0.018)
JPEN	2.517(0.214)	3.205(0.16)	1.182(0.084)	2.919(0.011)
Cov-I type covariance matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Ledoit-Wolf	33.2(0.04)	36.7(0.03)	36.2(0.03)	48.0(0.03)
Glasso	15.4(0.25)	16.1(0.4)	14.0(0.03)	14.9(0.02)
PDSCE	16.5(0.05)	16.33(0.04)	16.9(0.03)	17.5(0.02)
BLThresh	15.7(0.04)	17.1(0.03)	13.4(0.02)	17.5(0.02)
JPEN	7.1(0.042)	11.5(0.07)	8.4(0.042)	7.8(0.034)

The average relative error and their standard deviations (in percentage) for covariance matrix estimates are given in Table 8.1 and Table 8.2. The numbers in the bracket are the standard errors of relative error based on the estimates using different methods. Among all the methods JPEN and PDSCE perform similar for most of choices of n and p for all five type of covariance matrices. This is due to the fact that both PDSCE and JPEN use quadratic optimization function with a different penalty function. The behavior of Bickel and levina's estimator is quite good in Toepltiz case where it performs better than the other methods. For this type of covariance matrix, the entries away from the diagonal decay to zero and therefore soft-thresholding estimators like BLThresh perform better in this setting. However for neighorhood and hub type covariance matrix which are not necessarily banded type, Bickel and Levina estimator is not a natural choise as their estimator would fail to recover the underlying sparsity pattern. The performance of Ledoit-Wolf estimator is not very encouraging for Cov-I type matrix, This is beacuse Ledoit-Wolf estimator is generally not sparse and uniformly shrinks the sample covariance matrix towards identity matrix.

Figure 8.2: 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.



8.3 Recovery of Eigen-structure and Sparsity

To see the performance JPEN estimator in recovering the eigenstructure and sparsity, we plot the recovered eigenstructure and sparsity pattern in the both settings namely: (i) when the true covariance matrix has low condition number, and (ii) when the true covariance matrix has a very high condition number. The eigen-plots in Figure 8.3 and 8.4 show that among all the methods, estimates of eigenvalues of JPEN estimator are most consistent for the true eigenvalues. For Cov-I type covariance matrix where most of eigenvalues are close to zero and widely spread, the performance of JPEN estimator is impressive. This clearly shows the advantage of JPEN estimator of covariance matrix when the true eigenvalues are

dispersed or close to zero. The eigenvalues plot in Figure 8.4 shows that when eigen-spectrum of the true covariance matrix are not highly dispersed, the JPEN and PDSCE estimates of eigenvalues are almost the same, because both of these methods exploit Frobenius norm loss function with different penalty functions.

Figure 8.3: Eigenvalues plot for $n = 100,\ p = 50$ based on 50 realizations for neighborhood type of covariance matrix

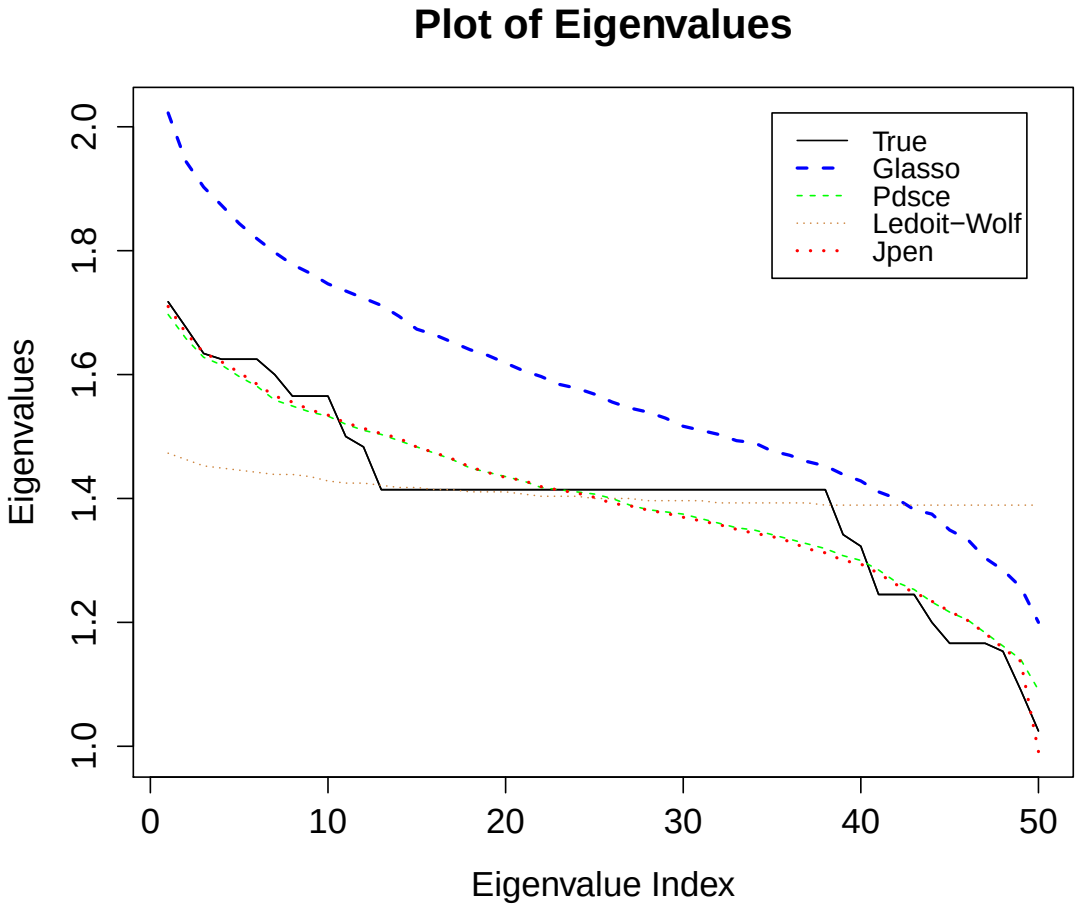
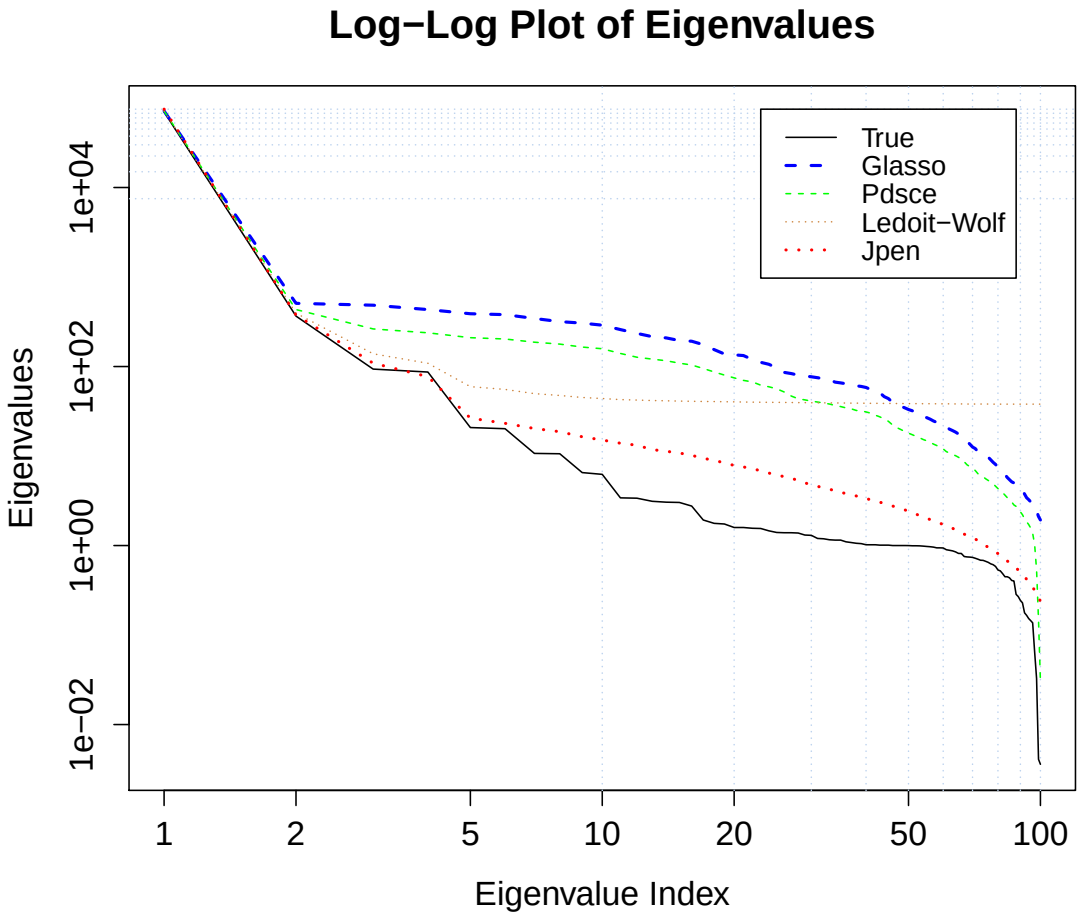


Figure 8.4: Eigenvalues plot for $n = 100$, $p = 100$ based on 50 realizations for Cov-I type matrix



Part II

Estimating Inverse Covariance Structure

CHAPTER 9

INVERSE COVARIANCE MATRIX AND ITS APPLICATIONS

9.1 Motivation

In many of the scientific applications the interest is to estimate the inverse of the covariance matrix (also called precision matrix). Precision matrix is widely used in a variety of applications including: i) linear discriminant classification: the classifier is function of precision matrix ii) Gaussian graphical modeling: A zero entry of the precision matrix implies conditional independence between the corresponding variables, iii) regression analysis: the regression coefficients are functions of precision matrix, and iv) Confidence interval estimation of population mean vector when the underlying data distribution is Gaussian.

In high dimensional data analysis problems where one often has very few observations as compared to the number of variables, the inverse of sample covariance matrix is not very useful. In fact for $n < p$ the sample covariance is singular and the inverse is not defined. In such situations one can replace the precision matrix by its generalized inverse [Rao [1972]]. A matrix G is called generalized inverse of A, if $AGA{=}A$. Generalized inverse have been used in number of applications including system of normal equations with singular matrix, least squares theory to express the variance of estimates. Although these are appealing in number of applications, their application in high dimensional analysis is often limited due to fact that they still remain singular and non-sparse. In particular their eigenvalues are biased, still remain over-dispersed compared to their population counterparts.

9.2 Related Work

Regularization is the most widely used technique to impose some structure on the estimation of precision matrices. In the likelihood based estimation framework, there is plenty of literature for the estimation of sparse precision matrix. Dempster [1972] introduced the concept of covariance selection where certain entries of precision matrices are set to zero, location of such entries in precision matrix is based on information in sample covarinace matrix. In likelihood framework, Gaussian distribution is most widely used data distribution for covariance matrix estimation due to its concavity as a function of precision matrix. In such framework, the optimization can be carried out by a number of fast algorithms, *viz.*, interior based methods (Vandenberghe and Boyd [2004],Vandenberghe et al. [1998]), subgradient based methods (Beck and Teboulle [2009]), and proximal gradient methods (Bertsekas [2010]). Among the early work on ℓ_1 regularized high dimensional covariance matrix estimation, Banerjee et al. [2008] proposed an estimator of sparse precision natrix (based on Σ) as the solution to the following optimization problem:

$$\hat{\Sigma}^{-1} = \arg\max_{\Omega} -log(det(\Omega)) + \frac{1}{2}tr(S\Omega) + \lambda\|\Omega\|_1, \tag{9.2.1}$$

where λ is non-negative tuning parameter that controls the level of sparsity in the estimate. They show that the above problem is convex and consider the estimation of Σ (rather than Σ^{-1}) and argue that one can solve the problem by optimizing over each row and corresponding column of $\Omega = \Sigma^1$ in a block coordinate descent fashion. In another interesting work, Meinshausen and Bühlmann [2006] take another approach by solving a lasso problem to each variable, using others as predictors. The component $\hat{\Sigma}^{-1}_{ij}$ is estimated to be zero if either the estimated coefficient of variable i on j, or the estimated coefficient of variable j on i, is non-zero (alternatively they use an AND rule). They show that asymptotically, this consistently estimates the set of non-zero elements of Σ^{-1}. However as argued in Friedman et al. [2008], this approach does not yield maximum likelihood estimator. Friedman et al. [2008] proposed Graphical lasso estimator as solution to (9.1) but instead they solve for Ω.

Using the fact that (9.1) is a smooth function of Ω except at the origin, they use subgradient based approach to solve a lasso least square regression of each variable, taking others as predictors. The graphical lasso uses cordinate descent algorithm and is very fast. Rothman et al. [2008] proposed an estimator (they call it SPICE) of precision matrix as solution to ℓ_1 regularized log likelihood function, however they only penalize off-diagonal entries. Other authors have proposed exact minimization of ℓ_1 regularized log-likelihood; Yuan and Lin [2007], Yuan [2009], Zhou et al. [2011],and Pourahmadi [2007, 2011]. In Cai et al. [2011], the authors proposed an estimator of precsion matrix as a solution to the following optimization problem.

$$\min_{\Omega} \|\Omega\|_1 \qquad \text{subject to} \qquad \|S\Omega - I\|_\infty \leq \lambda, \quad \Omega \in \mathbb{R}^{p\times p} \tag{9.2.2}$$

where λ is a tuning parameter. They gave a symmetric estimator based on this solution by taking

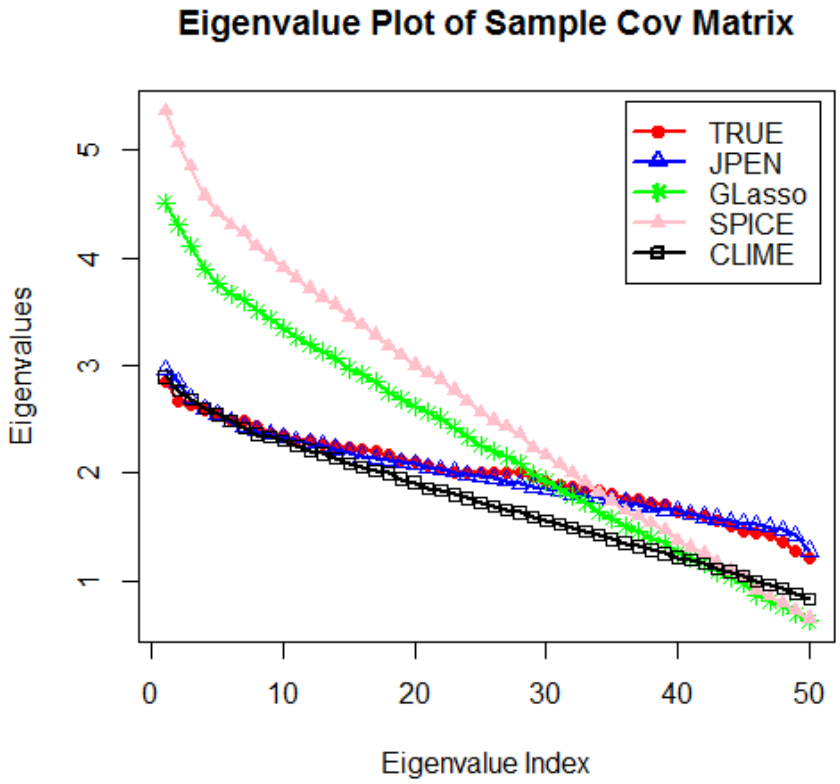
$$\hat{\omega}_{ij} = \hat{\omega}_{ij}\mathbf{1}_{\{|\hat{\omega}_{ij}|\leq|\hat{\omega}_{ji}|\}} + \hat{\omega}_{ji}\mathbf{1}_{\{|\hat{\omega}_{ji}|\leq|\hat{\omega}_{ij}|\}}$$

where $\hat{\omega}_{ij}$ is the $(i,j)^{th}$ entry of the solution $\hat{\Omega}$ of (9.2.2). Next, we describe the joint penalty estimator of precision matrix.

9.3 Joint Penalty for Precision Matrix Estimation

The ℓ_1 regularized precision matrix estimators often perform very good in estimating a sparse precision matrix, however whether they overcome the over-dispersion in eigenvalues, it is hard to justify. Figure 9.3.1 shows the eigenvalues of the estimated inverse precision matrix based on different methods. The methods consider here are: (i) Joint Penalty (JPEN) (see chapter 12 for JPEN estimator), (ii) Graphical Lasso (Glasso), (iii) SPICE, and (iv) CLIME.

Figure 9.1: Eigenvalues plot of precision matrix



Among these methods, the proposed Joint Penalty estimated eigenvalues are closest to the true eigenvalues, as it clearly reduces the over dispersion. This suggests that including a penalty on the variance of eigenvalues improves the over dispersion in the eigenvalues of precision matrix. Next we highlight few applications of precision matrix context of high dimensional data analysis.

9.4 Some Applications

9.4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is one of the widely used classification method in high dimensional setting. For simplicity, consider the two class classification problem. Given sample observations on class labels and features, LDA classifies each test observation x to either class k = 0 or k = 1 using the rule

$$\delta_k(x) = \arg\max_k \Big\{ x^T \hat{\Omega} \hat{\mu_k} \ - \ \frac{1}{2} \hat{\mu_k} \hat{\Omega} \hat{\mu_k} \ + \ log(\hat{\pi}_k) \Big\}, \tag{9.4.1}$$

where $\hat{\pi}_k$ is the proportion of class k observations in the training data, $\hat{\mu_k}$ is the sample mean for class k on the training data, and $\hat{\Omega} := \hat{\Sigma}^{-1}$ is an estimator of the inverse of the common covariance matrix.

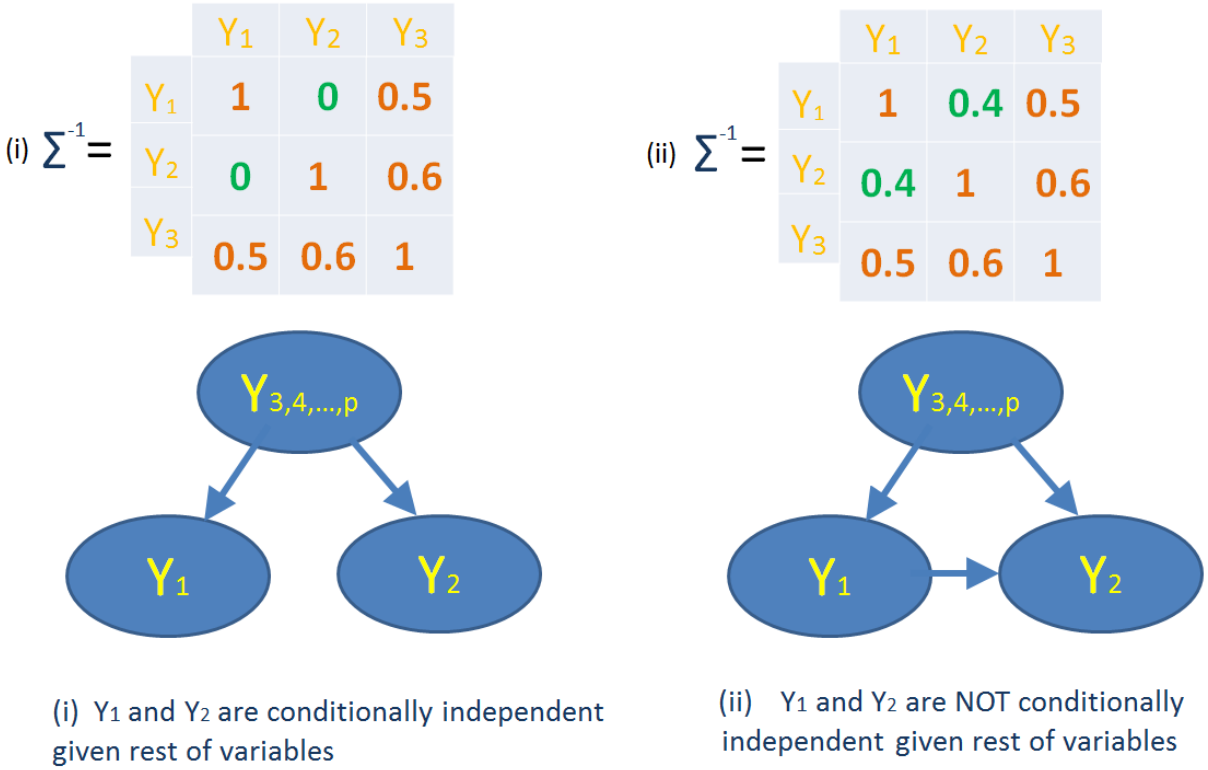
9.4.2 Gaussian Graphical Modeling

Given a random vector $Y = (Y_1, Y_2, \cdots, Y_p)$, Y_i, Y_j are said to be conditionally independent, if their joint distribution given rest of variables is same as the product of their conditionally marginal distribution. i.e.

$$P\Big(Y_i, Y_j \Big| \{Y \backslash Y_i, Y_j\}\Big) = P\Big(Y_i \Big| \{Y \backslash Y_i, Y_j\}\Big) P\Big(Y_j, \Big| \{Y \backslash Y_i, Y_j\}\Big),$$

where $\{Y \backslash Y_i, Y_j\}$ is vector Y excluding random variables Y_i and Y_j.

Figure 9.2: Illustration of conditional independence



In Gaussian graphical model, $Y = (Y_1, Y_2, \cdots, Y_p)$ follows multivariate normal distribution with mean vector μ and variance covariance matrix Σ. In this setting, the Y_i and Y_j are conditionally independent given the rest of variables $\{Y_k, k \neq i, k \neq j\}$, if partial correlation coefficient between Y_i and Y_j is zero which is equivalent to $\Sigma^{-1}_{ij} = 0$. This parallel relationship between the graph structure and the precision matrix greatly simplifies the estimation of graph structure to the problem of sparse precision matrix estimation.

CHAPTER 10

A JOINT CONVEX PENALTY(JCP) ESTIMATION

10.1 Motivation

Discussions from the above chapters 5 and 6 suggest estimation based on just ℓ_1 regularization does not yield appropriate shrinkage of the eigenspectrum. Therefore it is natural to impose an additional penalty to obtain a well-conditioned precision matrix estimation. In this chapter, we extend the likelihood based ℓ_1 regularized precision matrix estimation by adding trace norm penalty, and call it by Joint Convex Penalty (JCP) method. We implement **proximal gradient** method for computing the proposed estimator. The proposed algorithm is shown to converge at a rate $O(1/k)$ under mild conditions.

10.2 Joint Convex Penalty Estimation

The method to overcome the over-dispersion in precision matrix was previously studied by many authors with majority of work focusing on a well-conditioned estimation. Sheena and Gupta [2003] propose a constrained maximum likelihood estimator with restrictions on the lower or upper bound of the eigenvalues. This method focuses on only one of the two ends of the eigen-spectrum and thus the resulting estimator does not correct for the overestimation of the large eigenvalues and underestimation of the small eigenvalues simultaneously. Consequently their approach does not address the distortion of the entire eigen-spectrum – especially in high dimensional setting. Won et al. [2012] consider a maximum likelihood estimation of covariance matrix with condition number constraint. However this approach itself requires an estimation of condition number.

To control the distortion of eigen-spectrum of the covariance matrix, we consider a

joint penalty of sum of singular values (trace norm) in addition to ℓ_1 norm. By minimizing the joint penalty function of ℓ_1 and trace norm, the resulting estimated precision matrix is sparse as well as singular values of the corresponding covariance matrix are more centered than those of the sample observed covariance matrix.

10.2.1 Problem Formulation

Let $X \sim N_p(0, \Sigma)$, $\Sigma \succ 0$. For the simplicity of notation, we denote $\Omega = \Sigma^{-1}$. Let $\|\Omega\|_*$ be trace norm and defined as sum of singular values of the matrix Ω. Next we describe the proposed Joint Convex Penalty (JCP) estimator.

10.2.2 Proposed Estimator

We consider the following optimization problem with joint convex penalty:

$$\arg\min_{\Omega \succ 0} \quad F(\Omega) := f(\Omega) + g_1(\Omega) + g_2(\Omega). \tag{10.2.1}$$

where

$$f(\Omega) = -\log(det(\Omega)) + tr(S\Omega) \; ; \quad g_1(\Omega) = \lambda\|\Omega\|_1 \; ; \quad g_2(\Omega) = \tau\|\Omega\|_* \tag{10.2.2}$$

where λ and τ are non negative constants. The proposed algorithm to solve above problem is given in chapter 11 and can be used to solve a wide array of problems in statistics and machine learning. Some of the other important applications of this method include Matrix Classification Problems [Tomioka and Aihara [2007], Bach [2008], Multi-Task Learning Argyriou et al. [2008]].

Note that the $f(\Omega)$ in (10.2.2) 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 Fazel [2002]. The above problem is a convex optimization problem with non-smooth constraints. A natural choice to solve the above optimization problem is subgradient method

which generates a sequence of estimates $\{\Omega_k,\ k = 1, 2, 3, ...\}$ as

$$\Omega_k = \Omega_{k-1} - \alpha \bigtriangledown F(\Omega_{k-1}), \tag{10.2.3}$$

where α is some positive step size and $\bigtriangledown F(\Omega_{k-1})$ is sub-gradient indicating direction of greatest value increase of the function $F(\Omega)$ at Ω_{k-1}. This method has a well known convergence rate of $O(k^{-\frac{1}{2}})$ for non-smooth convex functions (Nesterov [2005]). We employ proximal gradient method to obtain a better rate of convergence of order $O(k^{-1})$. This method can be generalized to solve an arbitrary combination of convex functions [Bertsekas [2010]].

CHAPTER 11

PROXIMAL GRADIENT ALGORITHMS AND ITS CONVERGENCE ANALYSIS

11.1 Introduction

Much like Newton's method is a standard tool for solving unconstrained smooth optimization problems of modest size, proximal gradient algorithms can be viewed as an analogous tool for non-smooth, constrained, large-scale, or distributed versions of these problems. They are very generally applicable, but are especially well-suited to problems of substantial recent interest involving large or high-dimensional data sets. The main reason behind the success of proximal gradient algorithm is the availability of inexpensive operators of some well known functions, The projection gradient algorithm involves projecting a point onto a convex set, and often admits a closed form solution that can be obtained very quickly with a simple specialized methods. In our setup, we require computation of proximal operator for ℓ_1 and trace norm.

11.2 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,

$$\Omega_k = Prox_h(\Omega_{k-1}) = \arg\min_{\Omega \succ\ 0} \left(h(\Omega)\ +\ \frac{1}{2}\|\Omega - \Omega_{k-1}\|_2^2 \right). \tag{11.2.1}$$

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.

11.3 Basic Approximation Model

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

$$Q_L(\Omega,\Omega^{'}) := f(\Omega^{'}) + \ <\Omega - \Omega^{'}, \bigtriangledown f(\Omega^{'})> + \frac{L}{2}\|\Omega - \Omega^{'}\|_2^2 \tag{11.3.1}$$

where <A,B> is the inner product of A and B, and L is a positive constant. The optimization problem in (11.3.1) has two convex penalties. Proximal gradient method consists of sequential optimization of (11.3.1) by taking one constraint at a time in either cyclic or random order. Rewriting the optimization problem (11.3.1) with single constraint

$$\begin{aligned} Prox_{\frac{1}{L}g_i}(\Omega') &= \arg\min_{\Omega \succ 0} \left(Q_L(\Omega,\Omega') \ + \ g_i(\Omega) \right) \\ &= \arg\min_{\Omega \succ 0} \left(\frac{L}{2}\|\Omega - \{\Omega^{'} - \frac{1}{L} \bigtriangledown f(\Omega^{'})\}\|_2^2 + \ g_i(\Omega) \right). \end{aligned} \tag{11.3.2}$$

In general, L is unknown and it is estimated as an upper bound of Lipschitz parameter of $\bigtriangledown f(\Omega)$ [Bach et al. [2011]]. In other words L satisfies:

$$\| \bigtriangledown f(\Omega) - \bigtriangledown f(\Omega')\|_2 \leq L\|\Omega - \Omega^{'}\|_2 \qquad \forall\ \Omega, \Omega^{'} \in \mathbb{D}om(f). \tag{11.3.3}$$

A common method of generating value of L is to do a line search. For the optimization problem (11.3.2) we sequentially generate new estimates and increase the value of L by a factor $\gamma > 1$ until the following condition is met :

$$f_L(\Omega_k) \leq f(\Omega_{k-1}) + \ <\Omega_k - \Omega_{k-1}, \bigtriangledown f(\Omega_{k-1})> + \frac{L}{2}\|\Omega_k - \Omega_{k-1}\|_2^2, \tag{11.3.4}$$

where Ω_k is a solution at k^{th} iteration. In Lemma 11.3.1 and 11.3.2 below, we give proximal gradient operator for ℓ_1 and trace norm.

Lemma 11.3.1. *Let $M \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$. The proximal operator of $\|.\|_1$ with constant λ is given by*

$$Prox_{\lambda\|\cdot\|_1}(M) = \arg\min_{C \succ 0} \left(\lambda\|C\|_1 + \frac{1}{2}\|C - M\|_2^2 \right), \tag{11.3.5}$$

where

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

and abs(M) entriwise maximum function for matrix M.

Proof. Proof of the lemma is given in appendix. □

Lemma 11.3.2. *Let $M \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$ and $M = U\Sigma V^T$ be singular value decomposition of M where $U \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{r}}$ and $V \in \mathbb{R}^{\boldsymbol{r}\times\boldsymbol{n}}$ have orthogonal columns, Σ is the diagonal matrix of singular values of M and r is rank of matrix M. Then proximal operator of $\|.\|_*$ with constant τ is given by*

$$Prox_{\tau\|\cdot\|_*}(M) = \arg\min_{C\succ 0}\left(\tau\|C\|_* + \frac{1}{2}\|C - M\|_2^2\right), \tag{11.3.6}$$

where $Prox_{\tau\|\cdot\|_*}(M) = U\Sigma_\tau V^T$, Σ_τ is diagonal matrix with $((\Sigma_\tau))_{ii} = max(0, \Sigma_{ii} - \tau)$, $\tau > \min_{i\leq p}(\Sigma_{ii})$.

Proof. Proof of the lemma is given in appendix. □

For ℓ_1 and **t**race norm, the proximal operators are inexpensive to calculate. This results in efficient optimization of the objective function. The proximal operator for ℓ_1 is elementwise soft-thresholding operator. The proximal operator for trace norm is obtained by shrinking the singular values of precision matrix by regularization parameter. A larger value of regularization parameter results is more shrinkage of the eigenvalues.

11.4 Algorithm for optimization

Below we summarize the optimization algorithm for (11.3.2).

Initialize $L_0 = 1$, $\gamma > 1, \Omega_0 = diag(1/diag(S))$.

Iterate:

- Step 1: Set $\bar{L} = L_{k-1}$.

- Step 2: While $F(\Omega^*) > Q_{\bar{L}}((\Omega^*), \Omega_{k-1}) + g_1(\Omega^*) + g_2(\Omega^*)$

  (where $\Omega^* = \arg\min_{\Omega \succ 0} Q_{\bar{L}}(\Omega, \Omega_{k-1}) + g_1(\Omega) + g_2(\Omega)$)

  Set $\bar{L} = \gamma \bar{L}$.

- Step 3: Set $L_k = \bar{L}$, $Set\ Z_k = \Omega_k - \frac{1}{L_k} \bigtriangledown f(\Omega_k), \text{Set} Z_{k+1} = Prox_{(\tau/L_k)g_2}(Z_k), \text{Set}\Omega_{k+1} = Prox_{(\lambda/L_k)g_1}(Z_{k+1})$.

- Repeat until convergence.

11.5 Choosing the Regularization Parameter

The choice of regularization parameter is a challenging problem in high dimensional data analysis. Regularization has clear benefit in producing sparse solution as well reduces false discovery rate. A smaller value of λ accounts for a sparser structure of the precision matrix. Some of the methods for choosing regularization parameter include K-fold cross validation (KCV), stability approach to regularization selection (StARS) (Liu et al. [2010]), Akaike-Information Criteria (AIC) and Bayesian Information criteria (BIC). Experiments [Liu et al. [2010]] have shown that AIC and BIC methods tend to give poor performance for smaller sample sizes. Also K-fold cross validation tends to select smaller values of regularization parameter and results in higher false discovery rate. We follow StARS approach for estimating the regularization parameters λ and τ. Next we present the convergence analysis of the algorithm given in 11.4.

11.6 Convergence Analysis

We use the proposition 3.1 from Bertsekas [2010],

Lemma 11.6.1. *Let* $\{\Omega_n,\ L_n, n = 1, 2, ...\}$ *be the sequence generated by algorithm given in*

§11.4. Let $c > 0$ be a constant satisfying,

$$\begin{aligned} max\{\bigtriangledown\|f(\Omega)\|, \bigtriangledown\|g_j(\Omega)\|\} &\leq c \quad and \\ max\{f(\Omega_n) - f(Z_{n+j-1}) \ , \ g_j(\Omega_n) - g_j(Z_{n+j-1})\} &\leq c \ \|\Omega_n - Z_{n+j-1}\|, \quad j = 1, 2. \end{aligned}$$

Then for a cyclic order optimization of components $g_1(\cdot)$ and $g_2(\cdot)$, following holds :

$$\|\Omega_{n1} - \Omega^*\|^2 \leq \|\Omega_n - \Omega^*\|^2 \ - \ \frac{2}{L_n}(F(\Omega_n) - F(\Omega^*)) \ + \ 18c^2/L_n^2 \tag{11.6.1}$$

where $\Omega_{n1} = Prox_{(\tau/L_n)g_2}(\Omega_n)$ and Ω^* is a solution of (2.1) and (2.2).

Lemma 11.6.2. *Let $\{\Omega_n, \ L_n, n = 1, 2, ...\}$ be the sequence generated by algorithm §11.4. Let **c** be a constant as defined in Lemma (11.6.1). Then,*

$$F(\Omega_n) - F(\Omega^*) \leq \frac{L_n}{4}\left(\|\Omega_{n-1} - \Omega^*\|^2 - \|\Omega_n - \Omega^*\|^2\right) \ + \frac{9c^2}{2L_n} \ - \quad \frac{\lambda}{2} < \Omega^* - \Omega_n, \bigtriangledown\|\Omega_n\|_1 > \tag{11.6.2}$$

Proof. Proof of the Lemma 11.6.2 is given in appendix. □

We give below the convergence of the algorithm § 11.4.

Theorem 11.6.1. *Let $\{\Omega_k, k = 1, 2, ...\}$ be the sequence generated by algorithm §11.4. Let **c** be a constant as defined in Lemma 11.6.1. In addition, we assume that there exists a constant $M < \infty$ such that $\sum_{n=1}^{\infty} | < \Omega^* - \Omega_n, \bigtriangledown\|\Omega_n\|_1 > | \ < M$, then*

$$F(\Omega_k) - F(\Omega^*) \leq \left(\frac{\gamma L\|\Omega_0 - \Omega^*\|_F^2 \ + 18c^2 \ + \ M}{4k}\right), \tag{11.6.3}$$

where $L > 0$ is the least upper Lipschitz constant of the gradient of $f(\Omega)$ and $\gamma > 1$ is constant as defined in algorithm §11.4.

Proof. Note that $\frac{L_n}{\gamma} \leq L \leq L_n$, for all $n = 1, 2, ...$ Using Lemma 11.6.2, by adding $F(\Omega_n) - F(\Omega^*)$ over $n = 1, 2, ...$, we get the desired result. □

Due to the non-smoothness of the trace norm, the optimal first order black box methods have convergence rate of $O(k^{-\frac{1}{2}})$. The proximal gradient algorithm uses the special structure of the trace and ℓ_1 norm which improves the convergence rate for joint penalty to the order $O(k^{-1})$.

11.7 Simulation Study

To implement the proposed method, we perform a simulation study for various choices of precision matrix. We consider different types of precision matrices as outlined in chapter 8. Here we consider the underlying true precision matrix sparse rather than the covariance matrix. For all these choices of inverse covariance matrices, we generate random numbers from multivariate normal distribution with varying n and p. We set $n = 50, 100, 200$ and $p = 50, 100, 200$. The performance of proposed method is compared to graphical lasso and SPICE estimates of precision matrix. The joint convex penalty estimate of the precision matrix was computed using R software version 3.0.1 based on the algorithm §11.4. The graphical lasso estimate of the precision matrix was computed using R package "glasso" (http://statweb.stanford.edu/ tibs/glasso/). In "glasso" there is option of not penalizing the diagonal elements by setting the option "penalizing.diagonal=FALSE", this gives SPICE estimate.

11.7.1 Performance Criteria

For each of precision matrix estimate, we calculate Kullback-Leibler(KL) Loss, and Average Relative Error(ARE) defined below:

$$\begin{aligned} KLLoss(\Omega, \hat{\Omega}) &= -\log(det(\hat{\Omega})) + tr(\Omega^{-1}\hat{\Omega}) + \log(det(\Omega)) - p \\ ARE(\Omega, \hat{\Omega}) &= |\log(f(S, \hat{\Omega})) - \log(f(S, \Omega))| / \log(f(S, \Omega)) \end{aligned}$$

where $f(\cdot,\cdot)$ is density of multivariate Gaussian distribution and S is sample covariance matrix. The tuning parameters $\lambda$and τ were estimated following the Liu et al. [2010] criteria, which we describe below.

11.7.2 StARS Method of Tuning parameter selection:

Given a sample of size n, method generates N samples of size b, where $b < n$. In our setting for n<200, we choose b=0.8n and for n$\geq$ 200, b=$10\sqrt{n}$. For each of these N samples, an estimate of precision matrix is obtained. For each entry of the precision matrix, a measure of instability is calculated based on all N estimates. Finally a regularization parameter is selected which minimizes the average instability over all possible entries of the estimated precision matrix. In practice this method tends to select least amount of regularization parameter that simultaneously makes estimates of the precision matrix sparse and replicable under random sampling. StARS is used for estimating the penalization parameter for all the competing methods as given in simulation *viz.* JCP, Graphical Lasso and SPICE.

11.7.3 Simulation Results

The simulation results are given in tables 11.7.2.1-11.7.2.4. The numbers in bracket are standard error of the estimate based on 20 simulations. For Toeplitz type precision matrix, the proposed method outperforms other two in terms of MSE and KL-loss for small n. For large sample size, the Graphical lasso tends to give better performance than other methods. This may be due to fact that graphical lasso solves the constrained quadratic regression problem (dual of constrained objective function). For large sample size, the regression coefficients tends to approximate well the true values which results in better estimate of precision matrix.

11.7.3.1 Toeplitz Type Precision Matrix

Table 11.1: Average KL-Loss with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	**6**.15(0.065)	5.28(0.032)	4.829(0.0212)
Graphical Lasso	6.78(0.066)	**5**.17(0.054)	**4**.31(0.049)
SPICE	6.181(0.069)	5.28(0.037)	4.79(0.03)
p=100			
Mixed Penalty	**1**2.37(0.0806)	**1**0.84(0.0444)	9.922(0.0209)
Graphical Lasso	14.01(0.079)	11.21(0.061)	**8**.85(0.041)
SPICE	12.38(0.076)	11.08(0.043)	10.03(0.078)
p=200			
Mixed Penalty	**2**5.35(0.117)	**2**2.2(0.062)	20.39(0.021)
Graphical Lasso	31.5(0.124)	25.05(0.09)	**1**8.62(0.035)
SPICE	25.4(0.122)	22.25(0.07)	20.66(0.039)

Table 11.2: Average relative error with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	0.0857(0.0075)	0.0402(0.005)	0.1106(0.0041)
Graphical Lasso	0.135(0.01)	**0**.03(0.004)	**0**.0629(0.004)
SPICE	**0**.038(0.005)	0.067(0.005)	0.118(0.004)
p=100			
Mixed Penalty	0.0722(0.0054)	**0**.014(0.003)	0.1003(0.0031)
Graphical Lasso	0.24(0.006)	0.125(0.003)	**0**.024(0.003)
SPICE	**0**.031(0.01)	0.0802(0.004)	0.1235(0.004)
p=200			
Mixed Penalty	0.1182(0.0032)	**0**.009(0.0012)	0.09(0.001)
Graphical Lasso	0.456(0.003)	0.264(0.002)	**0**.04(0.0014)
SPICE	**0**.013(0.002)	0.1032(0.0032)	0.1313(0.0014)

11.7.3.2 Block Type Precision Matrix

Table 11.3: Average KL-Loss with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	**3**.624(0.067)	2.809(0.026)	2.46(0.0165)
Graphical Lasso	4.219(0.0652)	2.943(0.0321)	**2**.146(0.025)
SPICE	3.646(0.0644)	**2**.79(0.031)	2.37(0.0262)
p=100			
Mixed Penalty	**7**.53(0.081)	**6**.01(0.035)	5.232(0.0163)
Graphical Lasso	9.477(0.0717)	6.791(0.0485)	**4**.69(0.031)
SPICE	7.602(0.0856)	6.063(0.0413)	5.002(0.0372)
p=200			
Mixed Penalty	**1**5.05(0.123)	12.34(0.0443)	10.8(0.0264)
Graphical Lasso	22.46(0.2643)	16.29(0.0807)	**1**0.42(0.059)
SPICE	15.44(0.1296)	**1**2.27(0.056)	10.93(0.0444)

Table 11.4: Average relative error with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	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			
Mixed Penalty	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			
Mixed Penalty	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)

11.7.3.3 Hub Graph Type Precision Matrix

Table 11.5: Average KL-Loss with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	2.813(0.0618)	**1**.74(0.027)	**1**.066(0.0175)
Graphical Lasso	3.325(0.0479)	1.962(0.022)	1.098(0.0242)
SPICE	**2**.697(0.056)	1.89(0.0373)	1.399(0.0213)
p=100			
Mixed Penalty	6.576(0.115)	4.412(0.0545)	**2**.91(0.029)
Graphical Lasso	7.993(0.1087)	5.524(0.0693)	3.232(0.038)
SPICE	**5**.59(0.0792)	**4**.25(0.03)	3.46(0.0317)
p=200			
Mixed Penalty	10.06(0.1076)	5.862(0.0628)	3.163(0.0292)
Graphical Lasso	16.36(0.0978)	11.66(0.0585)	5.605(0.0358)
SPICE	**6**.88(0.098)	**4**.53(0.072)	**2**.83(0.0267)

Table 11.6: Average relative error with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	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			
Mixed Penalty	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			
Mixed Penalty	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)

11.7.3.4 Neighborhood Graph Type Precision Matrix

Table 11.7: Average KL-Loss with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	**3**.33(0.084)	**2**.28(0.036)	1.47(0.031)
Graphical Lasso	3.73(0.074)	2.45(0.045)	**1**.4(0.038)
SPICE	3.37(0.095)	2.723(0.0556)	2.274(0.0576)
p=100			
Mixed Penalty	6.507(0.165)	4.351(0.0472)	**2**.43(0.042)
Graphical Lasso	7.846(0.117)	5.472(0.0601)	2.598(0.0395)
SPICE	**5**.49(0.12)	**4**.21(0.038)	3.083(0.0576)
p=200			
Mixed Penalty	13.04(0.1767)	7.697(0.0877)	**4**.16(0.056)
Graphical Lasso	18.39(0.1129)	12.69(0.0815)	5.79(0.0639)
SPICE	**9**.3(0.131)	**6**.094(0.06)	4.359(0.0723)

Table 11.8: Average relative error with standard error over 20 replications

	n=50	**n=100**	**n=200**
p=50			
Mixed Penalty	0.0714(0.0029)	0.0323(0.0015)	**0**.008(0.001)
Graphical Lasso	0.068(0.0041)	0.0363(0.0022)	0.0128(0.0013)
SPICE	**0**.0056(0.001)	**0**.018(0.002)	0.0276(0.0014)
p=100			
Mixed Penalty	0.139(0.006)	0.069(0.0035)	0.0235(0.0005)
Graphical Lasso	0.184(0.005)	0.096(0.005)	0.0386(0.001)
SPICE	**0**.019(0.002)	**0**.005(0.001)	**0**.016(0.001)
p=200			
Mixed Penalty	0.2402(0.0029)	0.131(0.0013)	0.0428(0.0006)
Graphical Lasso	0.3591(0.0033)	0.2156(0.0018)	0.086(0.002)
SPICE	**0**.051(0.001)	**0**.01(0.003)	**0**.009(0.0004)

11.8 Summary

For block type precision matrix, the proposed method dominates the graphical lasso and SPICE in terms of KL-loss for small sample sizes. Also Joint Penalty and SPICE methods have better performance than graphical lasso in terms of average relative error. The corresponding standard error estimates for Joint Penalty and SPICE are smaller than those of graphical lasso. For neighborhood graph type precision matrix, the Joint Penalty method has better performance for small p in terms of KL-loss. However SPICE seems to perform better than other methods for small n and large p in terms of Average Relative Error.

Overall the Joint Penalty method has better performance than other the two methods for small sample size and for all choices of the precision matrix. The performance of all three methods methods varies over two choices of loss functions as they have different formulas. The value of KL-loss and average relative error are substantially different for different choices of the precision matrix. This shows that error in estimates depends upon the structure of underlying true precision matrix. For fixed p the estimates tends to improve for increasing sample sizes. However for fixed n, as expected, the estimators performance goes down with increasing p. For additional simulations, refer to Maurya [2014].

11.9 Discussion

The proposed method uses a joint penalty which is more flexible for penalizing the entries of precision matrix in different fashion than the Graphical Lasso and the SPICE. The proposed proximal gradient method can be extended to problems where one has an arbitrary number of convex penalty constraints. Under mild conditions the algorithm achieves sub-linear rate of convergence which makes it attractive choice for many optimization problems. The simulation study shows that the proposed method performs better than the other two methods for small sample sizes and large p. The simulated examples show that performance

of the JCP estimates of precision matrix are consistent as ARE and KL-Loss decreases rapidly (as well as the corresponding standard errors) for increasing sample size.

CHAPTER 12

SIMULTANEOUS ESTIMATION OF SPARSE AND WELL-CONDITIONED PRECISION MATRIX

In this chapter, we discuss the estimation of the precision matrix based on Frobenius norm loss function.

12.1 Motivation

Estimation of precision matrix based on the Frobenius norm loss function is not well defined problem unless S is well-conditined. One way to solve this problem is to first obtain a good estimator for the covariance matrix, then use this in place of S. We follow this approach and consider a two step procedure of precision matrix estimation. The main advantages of this approach are: (i) it allows to introduce sparsity in the precision matrix itself than in the covariance matrix, and (ii) the variance of eigenvalues penalty allows to perform optimal shrinkage in the eigenspectrum of the precision matrix. As the case in covariance matrix estimation, the resulting optimization problem is convex, which in trun yields an exact algorithm with low computational complexity. In this chapter we extend the JPEN approach to estimate a well-conditioned and sparse precision matrix. Similar to the covariance matrix estimation, we propose an estimator for the precision matrix based on regularized inverse correlation matrix and discuss its rate of convergence in both Frobenious and operator norm.

Notation: We shall use Z and Ω for inverse correlation and precision matrix respectively.

12.2 Joint Penalty Estimation: A Two Step Approach

Let $\hat{R}_K$ be a JPEN estimator for the true correlation matrix. By Lemma 6.3.2, $\hat{R}_K$ is positive definite and well-conditioned. Define the JPEN estimator of inverse correlation matrix as the solution to the following optimization problem,

$$\hat{Z}_K = \underset{Z=Z^T, tr(Z)=tr(\hat{R}_K^{-1})|(\lambda,\gamma)\in\hat{\mathscr{S}}_2^K}{\arg\min} \left[\|Z - \hat{R}_K^{-1}\|^2 \; + \; \lambda\|Z^-\|_1 \; + \; \gamma \sum_{i=1}^{p} \{\sigma_i(Z) - \bar{\sigma}(Z)\}^2 \right] \tag{12.2.1}$$

where

$$\hat{\mathscr{S}}_2^K = \Big\{ (\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\frac{\log\ p}{n}}, \forall \epsilon > 0, \\ \sigma_{min}\{(\hat{R}_K^{-1} + \gamma t_1 I) - \frac{\lambda}{2} * sign(\hat{R}_K^{-1} + \gamma t_1 I)\} > \epsilon \Big\},$$

and t_1 is the average of the diagonal elements of $\hat{R}_K^{-1}$. The minimization in (12.2.1) over Z is for fixed $(\lambda,\gamma) \in \hat{\mathscr{S}}_2^K$. The proposed JPEN estimator of the precision matrix (based on regularized inverse correlation matrix estimator $\hat{Z}_K$) is given by,

$$\hat{\Omega}_K = (S^+)^{-1/2} \hat{Z}_K (S^+)^{-1/2},$$

where S^+ is the diagonal matrix of the diagonal elements of S. Moreover (12.2.1) is a convex optimization problem and $\hat{Z}_K$ is positive definite.

Next we give another estimate of the precision matrix based on $\hat{\Sigma}_S$ of 6.3.3. Consider the following optimization problem:

$$\hat{\Omega}_S = \underset{\Omega=\Omega^T, tr(\Omega)=tr(\hat{\Sigma}_S^{-1})|(\lambda,\gamma)\in\hat{\mathscr{S}}_2^S}{\arg\min} \left[||\Omega - \hat{\Sigma}_S^{-1}||_F^2 + \lambda\|\Omega^-\|_1 + \gamma \sum_{i=1}^{p} \{\sigma_i(\Omega) - \bar{\sigma}_\Omega\}^2 \right], \tag{12.2.2}$$

where

$$\hat{\mathscr{S}}_2^S = \Big\{ (\lambda,\gamma) : \lambda,\gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\frac{\log\ p}{n}}, \; \forall \epsilon > 0, \\ \sigma_{min}\{(\hat{\Sigma}_S^{-1} + \gamma \; t_2 \; I) - \frac{\lambda}{2} * sign(\hat{\Sigma}_S^{-1} + \gamma t_2 I)\} > \epsilon \Big\},$$

and t_2 is average of the diagonal elements of $\hat{\Sigma}_S$. The minimization in (12.2.2) over Ω is for fixed $(\lambda, \gamma) \in \hat{\mathscr{S}}_2^S$. The estimator in (12.2.2) is positive definite and well-conditioned.

12.3 Weighted JPEN estimator for precision matrix

Similar to weighted JPEN covariance matrix estimator $\hat{\Sigma}_{K,A}$, a weighted JPEN estimator of the precision matrix is obtained by adding positive weights a_i to the term $(\sigma_i(Z)-1)^2$ in (12.2.2). The weighted JPEN precision matrix estimator is $\hat{\Omega}_{K,A} := (S^+)^{-1/2}\hat{Z}_A(S^+)^{-1/2}$, where

$$\hat{Z}_A = \underset{Z=Z^T, tr(Z)=tr(\hat{R}_K^{-1})|(\lambda,\gamma)\in\hat{\mathscr{S}}_2^{K,A}}{\arg\min} \quad \Big[||Z-\hat{R}_K^{-1}||_F^2 + \lambda\|Z^-\|_1 + \gamma\sum_{i=1}^{p} a_i\{\sigma_i(Z)-1\}^2\Big], \tag{12.3.1}$$

with

$$\hat{\mathscr{S}}_2^{K,A} = \Big\{(\lambda,\gamma) : \lambda \asymp \gamma \asymp \sqrt{\tfrac{\log\ p}{n}}, \lambda \le \tfrac{(2\ \sigma_{min}(R_K^{-1}))(1+\gamma t_1 max(A_{ii})^{-1})}{(1+\gamma\ min(A_{ii})^{-1}p} + \tfrac{\gamma min(A_{ii})}{p}\Big\},$$

and $A = \text{diag}(A_{11}, A_{22}, \cdots A_{pp})$ with $A_{ii} = a_i$. The optimization problem in (12.3.1) is convex and yields a positive definite estimator for $(\lambda, \gamma) \in \hat{\mathscr{S}}_2^{K,A}$.

12.4 Theoretical Analysis of JPEN estimators

In this section, we derive the rate of convergence of the proposed JPEN estimators of precision matrix in both Frobenius and spectral norm. Let $\Omega_0 = \Sigma_0^{-1}$ be the true precision matrix. Let $X = (X_1, X_2, \cdots, X_n)$ be sub-Gaussian random vectors as defined in (6.3.1). We make the following additional assumptions about the Ω_0. **B0.** Same as the assumption $A0$ of §6.3.

B1. With $H = \{(i,j) : \Omega_{0ij} \neq 0, i \neq j\}$, the $|H| \le s$, for some positive integer s.

B2. There exist $0 < \bar{k} < \infty$ large enough such that $(1/\bar{k}) \le \sigma_{min}(\Omega_0) \le \sigma_{max}(\Omega_0) \le \bar{k}$.

The next theorem gives consistency of $\hat{Z}_K$ and $\hat{\Omega}_K$.

Theorem 12.4.1. *Under Assumptions B0, B1, B2 and for* $(\lambda,\gamma) \in \hat{\mathscr{S}}_2^K$,

$$\|\hat{Z}_K - R_0^{-1}\|_F = O_P\Big(\sqrt{\frac{s\ \log\ p}{n}}\Big) \quad \textit{and} \quad \|\hat{\Omega}_K - \Omega_0\| = O_P\Big(\sqrt{\frac{(s+1)\ \log\ p}{n}}\Big) \tag{12.4.1}$$

where R_0^{-1} *is the inverse of true correlation matrix.*

Remark 12.4.1. Note that the JPEN estimator $\hat{\Omega}_K$ achieves mini-max rate of convergence for the class of covariance matrices satisfying assumption $B0$, $B1$, and $B2$ and therefore optimal. The similar rate is obtained in Cai et al. [2015] for their class of sparse inverse covariance matrices.

The next theorem gives consistency of $\hat{\Omega}_S$.

Theorem 12.4.2. *Let* $(\lambda,\gamma) \in \hat{\mathscr{S}}_2^S$. *Under Assumptions B0, B1, and B2, the* $\hat{\Omega}_S$ *of (12.2.1) satisfies,*

$$\|\hat{\Omega}_S - \Omega_0\|_F = O_P\Big(\sqrt{\frac{(s+p)\log\ p}{n}}\Big). \tag{12.4.2}$$

Consistency of $\hat{Z}_A$: A simple exercise shows that the estimator $\hat{Z}_A$ has similar rate of convergence as that of $\hat{Z}_K$.

CHAPTER 13

SIMULATIONS AND AN APPLICATION TO REAL DATA ANALYSIS

In this chapter, we compare the performance of the proposed method to various other methods for a number of structured precision matrices.

13.1 Simulation Results: Settings

We chose similar structures as in chapter 8 for precision matrix Ω_0 i.e. we replace Σ_0 by Ω_0 and for all these choices of inverse covariance matrices, generate random vectors from multivariate normal distributions with varying n and p. We chose $n = 50, 100$ and $p = 500, 1000$. The performance of proposed covariance matrix estimator $\hat{\Sigma}_K$ is compared to the following methods.

- **Graphical lasso [Friedman et al. [2008]]:** Graphical lasso estimates a sparse precision matrix. Here we invert the inverse, and include in our analysis. The estimate was computed using 'R' package 'Glasso'. For more details, refer to http://statweb.stanford.edu/tibs/glasso/.

- **Sparse Permutation Invariant Estimation (SPICE) [Rothman [2012]]:** the SPICE was computed using 'R' package 'Glasso' where we do not penalize diagonal elements. For more details, refer to (http://cran. r-project. org/web/ packages/PDSCE/index.html)

- **Constrained ℓ_1 minimization for inverse covariance matrices (CLIME) [Cai et al. [2011]]:** Their estimate was computed using 'R' package 'clime'. For more details see (https://cran.r-project.org/web/packages/clime/index.html).

All the computations were done using statistical software R on a AMD 2.8GHz processor.

Table 13.1: Precision matrix estimation

Block type precision matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Glasoo	4.144(0.523)	1.202(0.042)	0.168(0.136)	1.524(0.028)
PDSCE	1.355(0.497)	1.201(0.044)	0.516(0.196)	0.558(0.032)
CLIME	4.24(0.23)	6.56(0.25)	6.88(0.802)	10.64(0.822)
JPEN	1.248(0.33)	1.106(0.029)	0.562(0.183)	0.607(0.03)
Hub type precision matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Glasoo	1.122(0.082)	0.805(0.007)	0.07(0.038)	0.285(0.004)
PDSCE	0.717(0.108)	0.702(0.007)	0.358(0.046)	0.356(0.005)
CLIME	10.5(0.329)	10.6(0.219)	6.98(0.237)	10.8(0.243)
JPEN	0.684(0.051)	0.669(0.003)	0.34(0.024)	0.337(0.002)
Neighborhood type precision matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Glasoo	1.597(0.109)	0.879(0.013)	1.29(0.847)	0.428(0.007)
PDSCE	0.587(0.13)	0.736(0.014)	0.094(0.058)	0.288(0.01)
CLIME	10.5(0.535)	11.5(0.233)	10.5(0.563)	11.5(0.245)
JPEN	0.551(0.075)	0.691(0.008)	0.066(0.042)	0.201(0.007)

13.2 Performance Comparison

For each of the precision matrix estimate, we calculate Average Relative Error (ARE) based on 50 iterations based on the formula given in (8.2.1). The optimal values of tuning parameters were obtained over a grid of values by minimizing $5-$fold cross-validation as explained in §7.1.2. The JPEN estimator $\hat{\Omega}_K$ outperforms other methods for the most of the choices of n and p for all five types of the precision matrices. Additional simulations (not included here) show that for $n \approx p$, all the underlying methods perform similarly and the estimates of their eigenvalues are also well aligned with true values. However in high dimensional setting, for large p and small n, their performance is different as seen in simulations of Table 13.1 and Table 13.2.

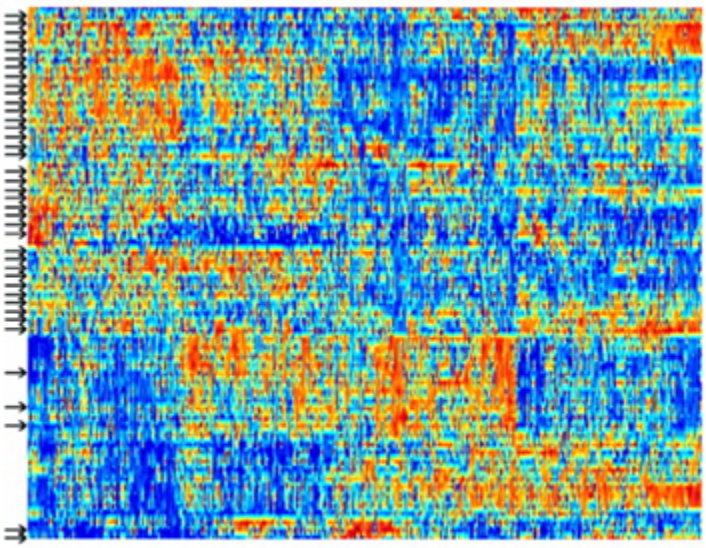
Table 13.2: Precision matrix estimation

Toeplitz type precision matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Glasoo	2.862(0.475)	2.89(0.048)	2.028(0.267)	2.073(0.078)
PDSCE	1.223(0.5)	1.238(0.065)	0.49(0.269)	0.473(0.061)
CLIME	4.91(0.22)	7.597(0.34)	5.27(1.14)	8.154(1.168)
JPEN	1.151(0.333)	2.718(0.032)	0.607(0.196)	2.569(0.057)
Cov-I type precision matrix				
	n=50		n=100	
	p=500	p=1000	p=500	p=1000
Glasoo	54.0(0.19)	190.(5.91)	14.7(0.37)	49.9(0.08)
PDSCE	28.8(0.19)	45.8(0.32)	16.9(0.04)	26.9(0.08)
CLIME	59.8(0.82)	207.5(3.44)	15.4(0.03)	53.7(0.69)
JPEN	26.3(0.36)	7.0(0.07)	15.7(0.08)	23.5(0.3)

13.3 Colon Tumor Gene Expression Data Analysis

In this section, we compare the performance of JPEN estimator of precision matrix for tumor classification using Linear Discriminant Analysis (LDA). The gene expression data (Alon et al. [1999] consists of 40 tumorous and 22 non-tumorous adenocarcinoma tissues. After preprocessing, data was reduced to a subset of 2,000 gene expression values with the largest minimal intensity over the 62 tissue samples (source: *http://genomics-pubs.princeton.edu/oncology /affydata/index.html*). Figure 13.1 shows the gene expression data for both tumorous and non-tumorous tissues. Tumorous tissues are marked with arrows on the left whereas normal tissues are unmarked. There is clear separation of normal and tumor tissues.

Figure 13.1: Colon tumor gene expression data



In our analysis, we reduced the number of genes by selecting p most significant genes based on ell_1 regularized logistic regression. We obtain estimates of precision matrix for $p = 50, 100, 200$ and then use the LDA to classify these tissues as either tumorous or non-tumorous (normal). Classify each test observation x to either class k = 0 or k = 1 using the LDA rule

$$\delta_k(x) = \arg\max_k \left\{ x^T \hat{\Omega} \hat{\mu_k} \; - \; \frac{1}{2} \hat{\mu_k} \hat{\Omega} \hat{\mu_k} \; + \; \log(\hat{\pi}_k) \right\}, \tag{13.3.1}$$

where $\hat{\pi}_k$ is the proportion of class k observations in the training data, $\hat{\mu}_k$ is the sample mean for class k on the training data, and $\hat{\Omega} := \hat{\Sigma}^{-1}$ is an estimator of the inverse of the common covariance matrix computed from the training data. Tuning parameters λ and γ were chosen using 5-fold cross validation. To create training and test sets, we randomly split the data into a training and test set of sizes 42 and 20, respectively; following the approach used by Wang et al. [2007], the training set has 27 tumorous tissues and 15 non-tumorous tissues. Since we do not have separate validation set, we do the 5-fold cross validation on

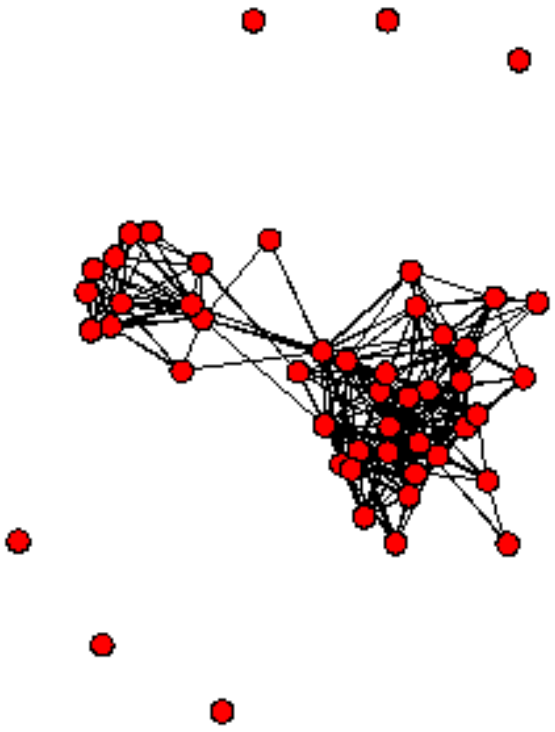
Table 13.3: Averages and standard errors of classification errors over 100 replications in %.

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)
Graphical Lasso	10.9(1.3)	9.4(0.89)	9.8(0.90)
Joint Penalty	**9.9(0.98)**	**8.9(0.93)**	**8.2(0.81)**

training data as following. At each split, we divide the training data into 5 subsets (fold) where 4 subsets are used to estimate the precision matrix and one subset is used to measure the classifier's performance, and for each split. This procedure is repeated 5 times by taking one of the 5 subsets as validation data. An optimal combination of λ and γ is obtained by minimizing the 5-fold cross validation error.

The average classification errors with standard errors over the 100 splits are presented in Table 13.3. Since the sample size is less than the number of genes, we omit the inverse sample covariance matrix as it is not well defined and instead include the naive Bayes', and support vector machine classifiers. Naive Bayes has been shown to perform better than the sample covariance matrix in high-dimensional settings (Bickel and Levina [2004]. Support Vector Machine (SVM) is another popular choice for high dimensional classification. Among all the methods in table 13.3, the precision matrix based LDA classifiers perform far better than Naive Bayes, SVM and Logistic Regression. For all other classifiers the classification performance deteriorates for increasing p. For larger p, i.e., when more genes are added to the data set, the performance of JPEN estimate based LDA classifier initially improves but it deteriorates for large p. For $p = 2000$, when all the genes are used in analysis, the classifier based on precision matrix has accuracy of 30%. This is due to the fact that as dimension of covariance matrix increases, the estimator does not remain very informative.

Figure 13.2: Partial correlation network of colon tumor gene expression data



To see the underlying gene-gene interaction in inverse correlation matrix, we do the network plot the genes that with non-zero partial correlation in the JPEN estimated inverse correlation matrices. The network graph shows the sparse structure underlying the colon tumor data, as there are few connected edges out of total 1225 possible edges.

APPENDIX

APPENDIX A

JPEN COVARIANCE MATRIX ESTIMATION

Proof. [Lemma 6.3.1]

Let

$$f(R) = \|R-K\|^2 + \lambda\|R^-\|_1 + \gamma\sum_{i=1}^{p}\{\sigma_i(R)-\bar{\sigma}_R\}^2. \tag{A.0.1}$$

where $\bar{\sigma}_R$ is the mean of eigenvalues of R. Due to the constraint $tr(R)=p$, we have $\bar{\sigma}_R = 1$. The third term of (A.0.1) can be written as

$$\sum_{i=1}^{p}\{\sigma_i(R)-\bar{\sigma}_R\}^2 = tr(R^2) - 2\ tr(R) + p$$

Then,

$$\begin{aligned} f(R) &= tr(R^2) - 2\ tr(RK) + tr(K^2) + \lambda\|R^-\|_1 + \gamma\{tr(R^2) - 2\ tr(R) + p\} \\ &= tr(R^2(1+\gamma)) - 2\ tr(K+\gamma\ I) + tr(K^2) + \lambda\|R^-\|_1 + p \\ &= (1+\gamma)\|R-(K+\gamma\ I)/(1+\gamma)\|^2 + tr(K^2) + \lambda\|R^-\|_1 + p \end{aligned} \tag{A.0.2}$$

This is quadratic in R with an ℓ_1 penalty to the off-diagonal entries of R, therefore a convex function in R. □

Proof. [Proof of Lemma 6.3.2] The solution to (A.0.2) satisfies:

$$2(R-(K+\gamma I))(1+\gamma)^{-1} + \lambda\frac{\partial\|R^-\|_1}{\partial R} = 0 \tag{A.0.3}$$

where $\frac{\partial\|R^-\|_1}{\partial R}$ is given by:

$$\frac{\partial\|R^-\|_1}{\partial R} = \begin{cases} 1 & : if \quad R_{ij} > 0 \\ -1 & : if \quad R_{ij} < 0 \\ \tau \in (-1,1) & : if \quad R_{ij} = 0 \end{cases}$$

Note that $\|R^-\|_1$ has the same value irrespective of sign of R, therefore the right hand side of (A.0.2) is minimum if :

$$\text{sign}(R) = \text{sign}(K + \gamma I) = \text{sign}(K)$$

for every $\epsilon > 0$, using (A.0.3), $\sigma_{min}\{(K+\gamma I) - \frac{\lambda}{2}\text{sign}(K)\} > \epsilon$ gives a $(\lambda, \gamma) \in \hat{\mathscr{S}}_1^K$ and such a choice of (λ, γ) guarantees the estimator to be positive definite.

Remark A.0.1. Intuitively, a larger γ shrinks the eigenvalues towards center which is 1, a larger γ would result in positive definite estimator, whereas a larger λ results in sparse estimate. A combination of (λ, γ) results in a sparse and well-conditioned estimator. In particular case, when K is diagonal matrix, the $\lambda < 2\gamma$.

□

Proof. [Theorem 6.3.1] Let $Q(R) = f(R) - f(R_0)$, where R_0 is the true correlation matrix and R is any other correlation matrix. Let $R = UDU^T$ be eigenvalue decomposition of R, where D is diagonal matrix of eigenvalues and U is matrix of eigen-vectors. $R_0 = U_0 D_0 U_0^T$ is eigenvalue decomposition of R_0. We have,

$$\begin{aligned} Q(R) = \|R - K\|_F^2 + \lambda\|R^-\|_1 + \gamma\ tr(D^2 - 2\ D + p) \\ - \|R_0 - K\|_F^2 - \lambda\|R_0^-\|_1 - \gamma\ tr(D_0^2 - 2\ D_0 + p) \end{aligned} \tag{A.0.4}$$

Let $\Theta_n(M) := \{\Delta : \Delta = \Delta^T,\ \|\Delta\|_2 = Mr_n,\ 0 < M < \infty\ \}$. The estimate $\hat{R}$ minimizes the $Q(R)$ or equivalently $\hat{\Delta} = \hat{R} - R_0$ minimizes the $G(\Delta) = Q(R_0 + \Delta)$. Note that $G(\Delta)$ is convex and if $\hat{\Delta}$ is its solution, then $G(\hat{\Delta}) \leq G(0) = 0$. Therefore, if we show that $G(\Delta)$ is non-negative for $\Delta \in \Theta_n(M)$, then $\hat{\Delta}$ will be within sphere of radius Mr_n. We require

$r_n = o\Big(\sqrt{(p+s)\ \log\ p/n}\Big)$. Consider,

$$\begin{aligned}
\|R-K\|_F^2 - \|R_0-K\|_F^2 &= tr(R^TR - 2R^TK + K^TK) - tr(R_0^TR_0 - 2R_0S + K^TK) \\
&= tr(R^TR - R_0^TR_0) - 2\ tr((R-R_0)^TK) \\
&= tr((R_0+\Delta)^T(R_0+\Delta) - R_0^TR_0) - 2\ tr(\Delta^TK) \\
&= tr(\Delta^T\Delta) - 2\ tr(\Delta^T(K-R_0))
\end{aligned}$$

Next, we bound term involving K in the above expression. We have

$$\begin{aligned}
|tr(\Delta^T(R_0-K))| &\leq \sum_{i\neq j} |\Delta_{ij}(R_{0ij}-K_{ij})| \\
&\leq \max_{i\neq j}(|R_{0ij}-K_{ij}|)\|\Delta^-\|_1 \\
&\leq C_0(1+\tau)\sqrt{\frac{\log\ p}{n}}\|\Delta^-\|_1 \leq C_1\sqrt{\frac{\log\ p}{n}}\|\Delta^-\|_1
\end{aligned}$$

holds with high probability, by a result (Lemma 1 from Ravikumar et al. [2011]) on the tail inequality for the sample covariance matrix of sub-Gaussian random vectors and where $C_1 = C_0(1+\tau), C_0 > 0$.

Next we obtain an upper bound on the terms involving γ in (A.0.4). By Cauchy-Schwarz inequality,

$$\begin{aligned}
&tr(D^2-2D) - tr(D_0^2-2D_0) \\
&= tr\{R^2-R_0^2\} - 2\ tr\{R-R_0)\} = tr(R_0+\Delta)^2 - tr(R_0^2) \\
&= 2\ tr(R_0\Delta) + tr(\Delta^T\Delta) \leq 2\ \sqrt{s}\|\Delta\|_F + \|\Delta\|_F^2.
\end{aligned}$$

To bound the term $\lambda(\|R^-\|_1 - \|R_0^-\|_1) = \lambda(\|\Delta^- + R_0^-\|_1 - \|R_0^-\|_1)$, let E be the index set as defined in Assumption A.2 of Theorem 6.3.1. Then using the triangle inequality, we obtain,

$$\begin{aligned}
\lambda(\|\Delta^- + R_0^-\|_1 - \|R_0^-\|_1) &= \lambda(\|\Delta_E^- + R_0^-\|_1 + \|\Delta_{\bar{E}}^-\|_1 - \|R_0\|_1) \\
&\geq \lambda(\|R_0^-\|_1 - \|\Delta_E^-\|_1 + \|\Delta_{\bar{E}}^-\|_1 - \|R_0^-\|_1) \\
&\geq \lambda(\|\Delta_{\bar{E}}^-\|_1 - \|\Delta_E^-\|_1)
\end{aligned}$$

Let $\lambda = (C_1/\epsilon)\sqrt{\log\ p/n}$, $\gamma = (C_1/\epsilon_1)\sqrt{\log\ p/n}$, where $(\lambda,\gamma) \in \hat{\mathscr{S}}_1^K$, we obtain,

$$\begin{aligned} G(\Delta) &\geq tr(\Delta^T\Delta)(1+\gamma) - 2\ C_1\Big\{\sqrt{\frac{\log\ p}{n}}(\|\Delta^-\|_1) + \frac{1}{\epsilon_1}\sqrt{\frac{s\ \log\ p}{n}}\|\Delta\|_F\Big\} \\ &\quad + \frac{C_1}{\epsilon}\sqrt{\frac{\log\ p}{n}}\big(\|\Delta^-_{\bar{E}}\|_1 - \Delta^-_E\|_1\big) \\ &\geq \|\Delta\|_F^2(1+\gamma) - 2C_1\sqrt{\frac{\log\ p}{n}}\big(\|\Delta^-_{\bar{E}}\|_1 + \|\Delta^-_E\|_1\big) \\ &\quad + \frac{C_1}{\epsilon}\sqrt{\frac{\log\ p}{n}}\big(\|\Delta^-_{\bar{E}}\|_1 - \Delta^-_E\|_1\big) - \frac{2C_1}{\epsilon_1}\sqrt{\frac{s\ \log\ p}{n}}\|\Delta\|_F. \end{aligned}$$

Also because $\|\Delta^-_E\|_1 = \sum_{(i,j)\in E, i\neq j}\Delta_{ij} \leq \sqrt{s}\|\Delta^-\|_F$,

$$\begin{aligned} -2C_1\sqrt{\frac{\log\ p}{n}}\|\Delta^-_{\bar{E}}\|_1 + \frac{C_1}{\epsilon}\sqrt{\frac{\log\ p}{n}}\|\Delta^-_{\bar{E}}\|_1 &\geq \sqrt{\frac{\log\ p}{n}}\|\Delta^-_{\bar{E}}\|_1\big(-2C_1 + \frac{C_1}{\epsilon}\big) \\ &\geq 0 \end{aligned}$$

for sufficiently small ϵ. Therefore,

$$\begin{aligned} G(\Delta) &\geq \|\Delta\|_F^2\big(1 + \frac{C_1}{\epsilon_1}\sqrt{\frac{\log\ p}{n}}\big) - C_1\sqrt{\frac{s\ \log\ p}{n}}\|\Delta^+\|_F\{1 + 1/\epsilon + 2/\epsilon_1\} \\ &\geq \|\Delta\|_F^2\Big[1 + \frac{C_1}{\epsilon_1}\sqrt{\frac{\log\ p}{n}} - \frac{C_1}{M}\{1 + 1/\epsilon + 2/\epsilon_1\}\Big] \\ &\geq 0, \end{aligned}$$

for all sufficiently large n and M. Which proves the first part of theorem. To prove the operator norm consistency, by sub-multiplicative norm property $\|AB\| \leq \|A\|\|B\|$,

$$\begin{aligned} \|\hat{\Sigma}_K - \Sigma_0\| &= \|\hat{W}\hat{R}\hat{W} - WKW\| \\ &\leq \|\hat{W} - W\|\|\hat{R} - K\|\|\hat{W} - W\| \\ &\quad + \|\hat{W} - W\|(\|\hat{R}\|\|W\| + \|\hat{W}\|\|K\|) + \|\hat{R} - K\|\|\hat{W}\|\|W\|. \end{aligned}$$

Since $\|K\| = O(1)$ and $\|\hat{R} - K\|_F = O(\sqrt{\frac{s\ \log\ p}{n}})$ these together implies that $\|\hat{R}\| = O_p(1)$. Also,

$$\begin{aligned} \|\hat{W}^2 - W^2\| &= \max_{\|x\|_2=1}\sum_{i=1}^p |(\hat{w}_i^2 - w_i^2)|x_i^2 \leq \max_{1\leq i\leq p}|(\hat{w}_i^2 - w_i^2)|\sum_{i=1}^p x_i^2 \\ &= \max_{1\leq i\leq p}|(\hat{w}_i^2 - w_i^2)| = O_p\big(\sqrt{\frac{\log\ p}{n}}\big), \end{aligned}$$

by using a result (Lemma 1 from Ravikumar et al. [2011]).

Next we shall shows that $\|\hat{W}-W\| \asymp \|\hat{W}^2-W^2\|$, (where A$\asymp$B means A=$O_P(B)$ and B=$O_P(A)$). We have,

$$\begin{aligned}\|\hat{W}-W\| &= \max_{\|x\|_2=1}\sum_{i=1}^{p}|(\hat{w}_i-w_i)|x_i^2 = \max_{\|x\|_2=1}\sum_{i=1}^{p}|\Big(\frac{\hat{w}_i^2-w_i^2}{\hat{w}_i+w_i}\Big)|x_i^2 \\ &\asymp \sum_{i=1}^{p}|(\hat{w}_i^2-w_i^2)|x_i^2 = C_3\|\hat{W}^2-W^2\|.\end{aligned}$$

where we have used the fact that the true standard deviations are well above zero, i.e., $\exists\ 0<C_3<\infty$ such that $1/C_3 \le w_i^{-1} \le C_3\ \forall\ i=1,2,\cdots,p$, and the sample standard deviations are all positive, i.e, $\hat{w}_i>0\ \forall\ i=1,2,\cdots,p$. Now since $\|\hat{W}^2-W^2\| \asymp \|\hat{W}-W\|$, it follows that $\|\hat{W}\|=O_p(1)$ and we have $\|\hat{\Sigma}_K-\Sigma_0\|^2 = O_p\Big(\frac{s\ \log\ p}{n}+\frac{\log\ p}{n}\Big)$. This completes the proof.

□

Proof. [Theorem 3.2] Let

$$f(\Sigma) = ||\Sigma-S||_F^2+\lambda\|\Sigma^-\|_1+\gamma\sum_{i=1}^{p}\{\sigma_i(\Sigma)-\bar{\sigma}_\Sigma\}^2,$$

Similar to the proof of theorem 6.3.1, define the function,

$$Q_1(\Sigma) = f(\Sigma)-f(\Sigma_0)$$

where Σ_0 is the true covariance matrix and Σ is any other covariance matrix. Let $\Sigma = UDU^T$ be the eigenvalue decomposition of Σ, where D is diagonal matrix of eigenvalues and U is matrix of eigen-vectors. Let $\Sigma_0 = U_0D_0U_0^T$ is eigenvalue decomposition of Σ_0. Then,

$$\begin{aligned}Q_1(\Sigma) = &\|\Sigma-S\|_F^2+\lambda\|\Sigma^-\|_1+\gamma\ tr(D^2)-(tr(D))^2/p \\ &-\|\Sigma_0-S\|_F^2-\lambda\|\Sigma_0^-\|_1-\gamma\ tr(D_0^2)-(tr(D_0))^2/p\end{aligned} \tag{A.0.5}$$

where $A = diag(a_1,a_2,\cdots,a_p)$. Write $\Delta=\Sigma-\Sigma_0$, and let $\Theta_n(M) := \{\Delta : \Delta=\Delta^T,\ \|\Delta\|_2 = Mr_n,\ 0<M<\infty\ \}$. The estimate $\hat{\Sigma}$ minimizes $Q(\Sigma)$ or equivalently $\hat{\Delta}=\hat{\Sigma}-\Sigma_0$ minimizes $G(\Delta)=Q(\Sigma_0+\Delta)$. Note that $G(\Delta)$ is convex and if $\hat{\Delta}$ be its solution, then we have

$G(\hat{\Delta}) \leq G(0) = 0$. Therefore if we can show that $G(\Delta)$ is non-negative for $\Delta \in \Theta_n(M)$, then $\hat{\Delta}$ will lie within the sphere of radius Mr_n. We require $\sqrt{(p+s)\ \log\ p} = o\Big(\sqrt{n}\Big)$.

$$\begin{aligned}
\|\Sigma - S\|_F^2 - \|\Sigma_0 - S\|_F^2 &= tr(\Sigma^T\Sigma - 2\Sigma^T S + S^T S) - tr(\Sigma_0^T\Sigma_0 - 2\Sigma_0 S + S^T S) \\
&= tr(\Sigma^T\Sigma - \Sigma_0^T\Sigma_0) - 2\ tr((\Sigma - \Sigma_0)S) \\
&= tr((\Sigma_0 + \Delta)^T(\Sigma_0 + \Delta) - \Sigma_0^T\Sigma_0) - 2\ tr(\Delta^T S) \\
&= tr(\Delta^T\Delta) - 2\ tr(\Delta^T(S - \Sigma_0))
\end{aligned}$$

Next, we bound the term involving S in the above expression, we have

$$\begin{aligned}
|tr(\Delta(\Sigma_0 - S))| &\leq \sum_{i\neq j} |\Delta_{ij}(\Sigma_{0ij} - S_{ij})| + \sum_{i=1} |\Delta_{ii}(\Sigma_{0ii} - S_{ii})| \\
&\leq \max_{i\neq j}(|\Sigma_{0ij} - S_{ij}|)\|\Delta^-\|_1 + \sqrt{p}\max_{i=1}(|\Sigma_{0ii} - S_{ii}|)\sqrt{\sum_{i=1}\Delta_{ii}^2} \\
&\leq C_0(1+\tau)\max_i(\Sigma_{0ii})\Big\{\sqrt{\frac{\log\ p}{n}}\|\Delta^-\|_1 + \sqrt{\frac{p\ \log\ p}{n}}\|\Delta^+\|_2\Big\} \\
&\leq C_1\Big\{\sqrt{\frac{\log\ p}{n}}\|\Delta^-\|_1 + \sqrt{\frac{p\ \log\ p}{n}}\|\Delta^+\|_2\Big\}
\end{aligned}$$

holds with high probability, by a result (Lemma 1 from Ravikumar et al. [2011]) where $C_1 = C_0(1+\tau)\max_i(\Sigma_{0ii}), C_0 > 0$ and Δ^+ is matrix Δ with all off-diagonal elements set equal to zero.

Next we obtain upper bound on the terms involving γ. we have,

$$tr(D^2) - (tr(D))^2/p - tr(D_0^2) - (tr(D))^2/p = tr(\Sigma^2) - tr(\Sigma_0^2) - (tr(\Sigma))^2/p + (tr(\Sigma_0))^2/p$$

(i)

$$\begin{aligned}
tr(\Sigma^2) - \Sigma_0^2)) &\leq tr(\Sigma_0 + \Delta)^2 - tr(\Sigma_0)^2 \\
&= tr(\Delta)^2 + 2\ tr(\Delta^2\Sigma_0) \leq tr(\Delta)^2 + C_1\sqrt{s}\|\Delta\|_F
\end{aligned}$$

(ii)

$$\begin{aligned}
tr((\Sigma))^2 - (tr(\Sigma_0))^2 &= (tr(\Sigma_0 + \Delta))^2 - (tr(\Sigma_0))^2 \\
&\leq (tr(\Delta))^2 + 2\ tr(\Sigma_0)\ tr(\Delta) \leq p\ \|\Delta\|_F^2 + 2\ \bar{k}p\sqrt{p}\|\Delta^+\|_F.
\end{aligned}$$

Therefore the γ term can be bounded by $2\|\Delta\|_F^2 + (C_1\sqrt{s} + 2\sqrt{\bar{p}}\bar{k})\|\Delta\|_F$. We bound the term involving λ as in similar to the proof of Theorem 6.3.1. For $\lambda \asymp \gamma \asymp \sqrt{\frac{\log\ p}{n}}$, the rest of the proof follows very similar to Theorem 6.3.1.

□

APPENDIX B

JCP FOR PRECISION MATRIX ESTIMATION

Proof. [Lemma 11.3.1] Let M^* be a solution of (11.3.6) which exits because (11.3.6) is a convex optimization problem. Then the following sub-gradient optimality condition holds [Vandenberghe and Boyd [2004]]

$$\mathbf{0} \in M^* - M + \lambda \partial \|M^*\|_1. \tag{B.0.1}$$

where $((\partial \|M\|_1))_{ij} = \partial |m_{ij}|$ and given by

$$\partial |m_{ij}| = \begin{cases} +1 & \text{if } \ m_{ij} > 0 \\ -1 & \text{if } \ m_{ij} < 0 \\ \in [-1,1] & \text{if } \ m_{i,j} = 0 \end{cases} \qquad i = 1,2,...m \ ; \quad j = 1,2....n.$$

Note that (B.0.1) is satisfied if and only if $|m_{ij}| \leq \lambda$ and therefore optimial solution is given by

$$m^*_{ij} = \text{sign}(m_{ij})(m_{ij} - \lambda)_+.$$

This completes the proof. □

Proof. [Lemma 11.3.2] Let L^* be a solution to (11.3.7). Then the following sub-gradient optimality condition holds :

$$0 \in L^* - M + \tau \partial \|L^*\|_* \tag{B.0.2}$$

Let $W = U \Sigma_\tau V^T$. We shall show that this choice of W satisfies the above optimality condition. The sub-differential $\partial \|W\|_*$ of $\|W\|_*$ is given by Bach [2008] as

$$\partial \|W\|_* = \left\{ UV^T + H \text{ such that } H \in \mathbb{R}^{\mathbf{m} \times \mathbf{n}}, \|H\|_2 \leq 1, U^T H = 0 \ \text{ and } \ HV = 0 \right\}.$$

Therefore

$$W - M + \tau\partial\|W\|_* = U\Sigma_\tau V^T - U\Sigma V^T + \tau(\ UV^T + H\ ).$$

Multiplying both sides by UU^T, and noting that $UU^T = I$ we obtain

$$\begin{aligned} W - M + \partial\|W\|_* &= UU^T(\ U\Sigma_\tau V^T - U\Sigma V^T + \tau(\ UV^T + H\ )\ ) \\ &= U\Sigma_\tau V^T - U(\Sigma - \tau I)V^T + UU^T H = 0. \end{aligned}$$

Therefore, $W = U\Sigma_\tau V^T$ is a solution to (11.3.7), this completes the proof. □

Proof. [Lemma 11.3.2] For such choice of L_n we have

$F(W_n) = f(W_n) + \lambda\|W_n\|_1 + \gamma\|W_n\|_* \ = Q_{L_n}(W_n, W_{n1}) + \lambda\|W_n\|_1 \ + \tau\|W_n\|_*$

$= f(W_{n1}) \ + \ \frac{L_n}{2}\|W_n - W_{n1}\|^2 + \ < W_n - W_{n1}, \bigtriangledown f(W_{n1}) > + \lambda\|W_n\|_1 \ + \tau\|W_n\|_*$

Also we have, $F(W^*) = \ f(W^*) + \lambda\|W^*\|_1 + \gamma\|W^*\|_*$

$f(W^*) \geq \ f(W_{n1}) + \ < W^* - W_{n1}, \bigtriangledown f(W_{n1}) >$

$\|W^*\|_1 \geq \ \|W_n\|_1 + \ < W^* - W_n, \bigtriangledown\|W_n\|_1 >$

$\|W^*\|_* \geq \ \|W_n\|_* + \ < W^* - W_n, \bigtriangledown\|W_n\|_* >$

We get ,

$$F(W^*) - F(W_n) \geq -\frac{L_n}{2}\|W_n - W_{n1}\|^2 + < W^* - W_n, \bigtriangledown f(W_{n1}) + \lambda \bigtriangledown \|W_n\|_1 + \tau \bigtriangledown \|W_n\|_* > \tag{B.0.3}$$

Note that W_n is solution of

$\bigtriangledown f(W_{n-1}) \ + L_n(W_n - W_{n1}) \ + \ \tau \bigtriangledown \|W_n\|_* = 0 \qquad$ (using 11.3.1)

Therefore (11.6.2) becomes

$$F(W^*) - F(W_n) \geq -\frac{L_n}{2}\left(\|W_{n1} - W_n\|^2 + 2 < W_{n1} - W_n, W_n - W^* > - \frac{2\lambda}{L_n} < W^* - W_n, \bigtriangledown\|W_n\|_1 >\right)$$

We know that for any three matrices $A,\ B,\ C$

$$\|B - A\|^2 + 2 < B - A, A - C > = \ \|B - C\|^2 - \|A - C\|^2$$

Using this, we obtain

$$F(W_n) - F(W^*) \leq \frac{L_n}{2}\left(\|W_{n1} - W^*\|^2 - \|W_n - W^*\|^2 - \frac{2\lambda}{L_n} < W^* - W_n, \bigtriangledown \|W_n\|_1 > \right).$$

Using lemma (11.3.2), we get

$$F(W_n) - F(W^*) \leq \frac{L_n}{4}\left(\|W_{n-1} - W^*\|^2 - \|W_n - W^*\|^2 \right) + \frac{9c^2}{2L_n} - \frac{\lambda}{2} < W^* - W_n, \bigtriangledown \|W_n\|_1 > .$$

This completes the proof. □

APPENDIX C

JPEN FOR PRECISION MATRIX ESTIMATION

Proof. [Theorem 12.4.1] To bound the cross product term involving Δ and $\hat{R}_K^{-1}$, we have,

$$\begin{aligned} |tr((R_0^{-1} - \hat{R}_K^{-1})\Delta)| &= |tr(R_0^{-1}(\hat{R}_K - R_0)\hat{R}_K^{-1}\Delta)| \\ &\leq \sigma_1(R_0^{-1})|tr((\hat{R}_K - R_0)\hat{R}_K^{-1}\Delta)| \\ &\leq \bar{k}\sigma_1(\hat{R}_K^{-1})|tr((\hat{R}_K - R_0)\Delta)| \\ &\leq \bar{k}\bar{k}_1|tr((\hat{R}_K - R_0)\Delta)|. \end{aligned}$$

where $\sigma_{min}(\hat{R}_K) \geq (1/\bar{k}_1) > 0$, is a positive lower bound on the eigenvalues of JPEN estimate $\hat{R}_K$ of the correlation matrix R_0. Such a constant exists by Lemma 6.3.2. Rest of the proof closely follows that of Theorem 6.3.1. □

Proof. [Theorem 12.4.2] We bound the term $tr((\hat{\Omega}_S - \Omega_0)\Delta)$ similar to that in proof of Theorem 12.4.1. Rest of the proof closely follows to that Theorem 12.4.1. □

BIBLIOGRAPHY

BIBLIOGRAPHY

I. Johnstone and Y. Lu. Sparse principal components analysis. *Unpublished Manuscript*, 2004.

H. Zou, Hastie T., and Tibshirani R. Sparse principal components analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.

K. Mardia, Kent J., and Bibby J. *Multivariate Analysis.*, volume 1. Academic Press, New York, NY, 1979.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

M. Wainwright, Ravikumar P., and Lafferty J. High-dimensional graphical model selection using l_1-regularized logistic regression. *Proceedings of Advances in Neural In formation Processing Systems.*, 2006.

M. Wainright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programmming (lasso). *IEEE Transactions on Information Theory archive*, 55, 2009.

M. Yuan. Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2009.

Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

D. Yatsenko, K. Josic, r A. S. Ecke, E. Froudarakis, R. J. Cotton, and A. S Tolias. Improved estimation and interpretation of correlations in neural circuits. *PLoS Comput. Biol*, 11, 2015.

V. Marcenko and L. Pastur. Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb*, 1:507–536, 1967.

S. Geman. A limit theorem for the norm of random matrices. *The Annals of Statistics*, 8(2): 252–261, 1980.

D.L. Donoho, M. Gavish, and I.M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *http://arxiv.org/pdf/1311.0851.pdf*, 2015.

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.

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.

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.

J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software*, 23:501–520, 2008.

A. Dempster. Covariance selection. *Biometrika*, 32:95–108, 1972.

T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38:2118–2144, 2010.

N. Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics*, 36:2717–2756, 2008.

A. Rothman, Bickel P. J., Levina E., and Zhu J. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

R. Tibshiran. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, pages 267–288, 1996.

S. Chaudhuri, M. Drton, and T.S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199–216, 2007.

A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, 27:12182–12186, 2000.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, 2009.

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.*, 34:1373–1403, 1986.

O. Ledoit and M. Wolf. Optimal estimation of a large-dimensional covariance matrix under stein's loss. *http://papers.ssrn.com/*, 2014.

Y. Sheena and A. Gupta. Estimation of the multivariate normal covariance matrix under some restrictions. *Statistics and Decisions*, 21:327–342, 2003.

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.

L. R. Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597, 1980.

S. Lin and M. D. Perlman. A monte carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, 6:411–429, 1985.

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

B. Rajaratnam, D. Vincenzi, and B. Naul. A theoretical study of stein's covariance estimator. *Technical report, Department of Statistics, Stanford University*, 2014.

A. Maurya. A sparse and well-conditioned estimation of covariance and inverse covariance matrices using a joint penalty. *Journal of Machine Learning Research*, 15-345, 2016.

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.

D.P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization, a survey. *Labaratory for Information and Decision Systems Report LIDS-P-2848. MIT*, 2010.

L. Vandenberghe and S. Boyd. Convex optimization. *Cambridge University Press*, 2004.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *In Optimization for Machine Learning, MIT press*, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Science*, 2:183–202, 2009.

J. Friedman, Hastie T., and Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics.*, 9(3):432–441, 2008.

C.R. Rao. Generalized inverse of a matrix and its applications. *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Univ. of Calif. Press*, 1:601–620, 1972.

L. Vandenberghe, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1998.

O. Banerjee, E.L. Ghaoui, and d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

S. Zhou, P. Rutimann, and Buẗhlmann P. Xu M. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 2011.

M. Pourahmadi. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika 94*, 4:1006–1013, 2007.

M. Pourahmadi. Modeling covariance matrices: The glm and regularization perspectives. *Statistical Science*, 26:369–387, 2011.

T. Cai, W. Liu, and X. Luo. , a constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, 106:2594–607, 2011.

R. Tomioka and K. Aihara. Classifying matrices with a spectral regularization. *Proc. 24th Int. Conf. Machine Learning*, pages 895–902, 2007.

F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

A. Argyriou, T. Evgeniou T., and M. Pontil. Convex multi-task feature learning. *Machine Learning, Special Issue on Inductive Transfer Learning*, 73:243–272, 2008.

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

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, pages 127–152, 2005.

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

H. Liu, K. Roede, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *In Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

A. Maurya. A joint convex penalty for inverse covariance matrix estimation. *Computational Statistics and Data Analysis*, 75:15–27, 2014.

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):6745–6750, 1999.

L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification. *Proceedings of the 24th International Conference on Machine Learning.*, pages 983–990, 2007.

P. Bickel and E. Levina. Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004.

P. Ravikumar, Wainwright M.and Raskutti G., and Yu B. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.