Title
Covariance matrix estimation and variable selection in high dimension

Author
Cai, Mu

Publication Date
2013

Peer reviewed|Thesis/dissertation

Covariance matrix estimation and variable selection in high dimension

by

Mu Cai

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel , Chair
Professor Cari Kaufman
Professor Noureddine El Karoui
Professor James L. Powell

Spring 2013

Covariance matrix estimation and variable selection in high dimension

Abstract

Covariance matrix estimation and variable selection in high dimension

by

Mu Cai

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter J. Bickel , Chair

First part of the thesis focuses on sparse covariance matrices estimation under the scenario of large dimension p and small sample size n. In particular, we consider a class of covariance matrices which are approximately block diagonal under unknown permutations. We propose a block recovery estimator and show it achieves minimax optimal convergence rate for the class, which is the same as if the permutation were known. The problem is also related to sparse PCA and k-densest subgraphs, where the spike model is a special case of their intersection. Simulations of the spike model and multiple block model, together with a real world application, confirm that the proposed estimator is both statistically and computationally efficient.

Second part of the thesis focuses on variable selection in linear regression, also under the high dimensional scenario of large p and small n. We propose a general framework to search variables based on their covariance structures, with a specific variable selection algorithm called $kForward$ which iteratively fits local/small linear models among relatively highly correlated variables. For simulation experiments and a real world data set, we compare $kForward$ to other popular methods including the $Lasso$, $ElasticNet$, $SCAD$, $MC+$, $FoBa$ for both variable selection and prediction.

Dedicated to my family and in memory of my grandmother Jia Zhen Yin.

Contents

Acknowledgments

It is a great pleasure to thank the many people who made this thesis possible. First and foremost I would like to thank my advisor Professor Peter Bickel. I am extremely grateful to Peter for his continuous support, guidance and insightful advice throughout the years. I am also very thankful to my committee, Professor Cari Kaufman, Professor Noureddine El Karoui and Professor James Powell, for their kind assistance and advice for my qualification exam and my thesis. I want to thank Professor Boaz Nadler from Weizmann Institute of Science and Professor Ya'acov Ritov from The Hebrew University for sharing interesting ideas and examples with me, many of which lead to the very foundation of my thesis. I also thank Dr. Aiyou Chen from Google for all the fruitful discussions which in many ways inspired this thesis. I am indebted to my colleagues Jingyi Jessica Li, Rachel Wang and Nathan Boley for sharing valuable data and application with me. I am grateful to Professor Jim Pitman, Professor Terry Speed, Professor Bin Yu, Professor Haiyan Huang, Dr. Choongsoon Bae, Dr. Sheng Ma, Professor Ching-Shui Cheng, Professor Yun Song, Ryan Lovett, La Shana Porlaris, Denise Yee, Judith Foster and many others for their kind advice and help. I also want to thank my undergraduate mentor Professor Almut Burchard, Professor Charles Pugh, Professor Steve Tanny and Professor B*á*lint Vir*á*g from University of Toronto. At last, I thank my family and all my friends, especially my parents Chong Bang Cai and Qiang Li and my fianc*é* Yi Ru Qin, for their endless support and encouragement.

Chapter 1

Notations

- Define $[n] = \{1, 2, ..., n\}$.
- For $L = \{l_1, l_2, ..., l_n\}$, denote $L_i = l_i$ for $i = 1, ..., n$.
- Denote $I_{n\times n}$ the n by n identity matrix.
- Denote $\mathcal{S}_p^+$ the set of $p \times p$ positive definite matrices.
- Denote Π_p the set of all permutations of $[p]$.
- For vector $v \in \mathbb{R}^n$,
  - l_q norm for $q \geq 1$: $\|v\|_q = (\sum_{i=1}^n |v_i|^q)^{\frac{1}{q}}$
  - l_0 norm: $\|v\|_0 = \sum_{i=1}^n \mathbb{I}(v_i \neq 0)$
  - l_∞ norm: $\|v\|_\infty = \max_{i=1}^n |v_i|$
  - $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$
  - $v_L = (v_i)_{i\in L} \in \mathbb{R}^{|L|}$ for $\forall L \subset [n]$
  - The support of v: $supp(v) = \{i : \mathbb{I}(v_i \neq 0)\}$
- For matrix $A \in \mathbb{R}^{n\times p}$, vector $v \in \mathbb{R}^n$,
  - $\lambda_{\max}(A)$: largest eigenvalue of A
  - $\lambda_{\min}(A)$: minimal eigenvalue of A
  - l_q operator norm for $q \geq 1$: $\|A\|_q = \max_{v\neq 0} \frac{\|Av\|_q}{\|v\|_q}$
  - Spectral norm for square A with $n = p$: $\|A\| = \|A\|_2 = \sqrt{\lambda_{\max}(A^tA)}$
  - l_1 norm: $\|A\|_1 = \max_{1\leq j\leq p} \sum_{i=1}^n |A_{ij}|$
  - l_∞ norm: $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^p |A_{ij}|$

- Frobenius norm: $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p |A_{ij}|^2}$
- $A_j \in \mathbb{R}^{n\times 1}$ is the jth column of A
- $A_{(i)} \in \mathbb{R}^{1\times p}$ is the ith row of A
- For $i_0 \in [n], j_0 \in [p], L \subset [n], J \subset [p], B \subset [p] \times [p]$,
  - $A_{L,J} = (A_{ij})_{i\in L, j\in J} \in \mathbb{R}^{|L|\times|J|}$
  - $A_J = (A_{ij})_{1\le i\le n, j\in J} \in \mathbb{R}^{n\times|J|}$
  - $A_{(L)} = (A_{ij})_{i\in L, 1\le j\le p} \in \mathbb{R}^{|L|\times p}$
  - $A_{i_0 J} = (A_{i_0 j})_{j\in J} \in \mathbb{R}^{1\times|J|}$
  - $A_{L j_0} = (A_{i j_0})_{i\in L} \in \mathbb{R}^{|L|\times 1}$
  - $A_B = (A_{ij}\mathbb{I}((i,j)\in B))_{1\le i,j\le p}$
  - $v^t A_B v = \sum_{(i,j)\in B}^p v_i v_j A_{ij}$
- $\bar{A} = (\bar{A}_j)_{1\le j\le p} \in \mathbb{R}^{1\times p}$.

Part I

High dimensional covariance matrix estimation

Chapter 2

Introduction

Covariance estimation plays a central role in many statistical methodologies including regression analysis, principal component analysis (PCA), linear and quadratic discriminant analysis (LDA, QDA). Suppose $X \in \mathbb{R}^{n\times p}$ is observed. The rows $X_{(1)}, ..., X_{(n)} \in \mathbb{R}^{1\times p}$ are i.i.d. p-variate random variables with covariance matrix $\Sigma = (\Sigma_{ij})_{1\leq i,j\leq p}$. The goal is to construct an estimator $\tilde{\Sigma}$ that is close to the population Σ. There are many metrics for measurements. For example element-wise estimation corresponds to minimizing $\max_{i,j} |\tilde{\Sigma}_{ij} - \Sigma_{ij}|$, while techniques like PCA and LDA require estimation of eigenvalues, eigenvectors of Σ, and measurement of $\tilde{\Sigma} - \Sigma$ with errors measured by the Frobenius norm $\|.\|_F$ or spectral norm $\|.\|$. A classical approach is to estimate Σ by the empirical covariance matrix $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)} - \bar{X})^T(X_{(i)} - \bar{X})$. Nowadays many applications involve high dimensional data with $p > n$ or $p = O(n)$. This poses many new challenges to classical statistics and extensive research has been done in this area. As an example, although $\hat{\Sigma}$ still performs well for element-wise estimation with a convergence rate of $\sqrt{\frac{1}{n}}$, it is well known it does not work for estimation of eigenvalues, eigenvectors for matrices Σ such as the identity, which are of high rank, and not approximable in the spectral norm by matrices of bounded rank.

However following work by various authors, Bickel and Levina (BL2008a-1) (BL2008b-1), T. Cai et al (CZZ2010-1) (CZ2012-1) , El Karoui (EK2007-1), it has been shown that if population matrices can be approximated in the spectral norm by structured matrices with sparse structure, then estimates of these matrices converging at reasonable rates can be constructed even if $p > n$. Such matrices arise naturally if there is a metric on the variables such that high distance between the variables is associated with local covariance. For instance, if $X_{(1)} = (X_{1t_1}, ..., X_{1t_p})$, where the t_j correspond to times in a given year, such a metric is the usual Euclidean metric $\rho(X_{1t_a}, X_{1t_b}) = |t_a - t_b|$. More complicated structures can arise from spatial fields. The natural approximation here are $\tilde{\Sigma}$ of the form:

$$\tilde{\Sigma}(\delta) = \Sigma_{ij}\mathbb{I}(\rho(X_{1i}, X_{1j}) \leq \delta) \tag{2.1}$$

Another possibility, see Furrer and Bengtsson (FB2007-1), is to replace the indicators by a

monotone tapering function $\sigma(\rho(X_{1t_a}, X_{1t_b}))$ which is monotone decreasing to 0 and

$$(\sigma(\rho(X_{1t_a}, X_{1t_b})))_{1\leq a,b\leq p}$$

is a positive definite matrix.

In examples such as we have given above, the metric is known and the approximating matrices can be readily constructed. However suppose there is a reason to believe that a metric of this type is present but not known in advance. For instance, suppose the variables are expression of genes in a biochemical pathway. The metric which is unknown, is roughly geodesic distance in the graph representing the pathway. We are then faced with approximating a given covariance matrix by a matrix which, after an unknown permutation of the variables, is of the given structure. That is the topic of the first part of the thesis.

We note that similar problems in which we assume that the approximating class has restrictions only on the member of zeros of the matrix, but not their position have been treated. El Karoui (EK2007-1) proposed and analyzed a class of covariance matrices with β-sparsity, which requires the number of walks of length k on the graph with adjacency matrix induced by the population covariance matrix is bounded by $O(p^{\beta(k-1)+1})$. He proposed an entry-wise thresholding estimator and showed that it is consistent in operator norm. Bickel and Levina (BL2008a-1) (BL2008b-1) proposed and analyzed approximately bandable class and classes with l_q ball constraint on each row. They also provided upper bound on spectral norm for corresponding thresholding and banding estimator. Another focus has been on approximating covariance matrices for particular purpose such as PCA, CCA, some of which assume structured sparse approximation, Amini and Wainwright (AW2009-1), Johnstone (J2001-1), and others does not, d'Aspermont et al(DA2007-1), Zou, Hastie and Tibshirani (ZHT2006-1), Joliffe et al (J2003-1). Another direction involves estimation of Σ^{-1}. For structured situations, the results of Bickel and Lindner (BL2010-1) suggest quite generally that inverting estimates of Σ taking advantage of the assumed structure works as well as possible. Also Bickel and Levina (BL2008a-1) (BL2008b-1) give methods for estimating Σ^{-1} directly in structurally approximable cases, and see graphical Lasso Friedman, Hastie and Tibshirani (FHT2008-1), Rothman et al (R2008-1) for unstructured sparse cases. We begin by reviewing some of this work.

2.1 Related Work

Following Bickel and Levina (BL2008a-1) (BL2008b-1), define classes of l_q sparse covariance matrices as:

$$\mathcal{U}_t(q, k_0, M_0) = \{\Sigma : \Sigma_{ii} < M_0, \sum_{j=1}^{p} |\Sigma_{ij}|^q \leq k_0, \text{ for } \forall i\} \tag{2.2}$$

The classes of approximately bandable covariance matrices is defined as:

$$\mathcal{U}_b(\alpha, C, M) = \{\Sigma : \max_j \sum_i \{|\Sigma_{ij}| : |i-j| > k\} \leq Ck^{-\alpha}, \lambda_{\max}(\Sigma) < M\} \tag{2.3}$$

Define thresholding estimator:

$$T_h(\hat{\Sigma}) = (\hat{\Sigma}_{ij}\mathbb{I}(|\hat{\Sigma}_{ij}| > h))_{1\le i,j\le p} \tag{2.4}$$

Define banded estimator:

$$B_k(\hat{\Sigma}) = (\hat{\Sigma}_{ij}\mathbb{I}(|i-j| \le k))_{1\le i,j\le p} \tag{2.5}$$

In a series of work by T. Cai et al (CZZ2010-1) (CL2011-1) (CY2012-1) (CZ2012-1) , minimax convergence rates in various norms and adaptive estimators were developed for similar classes of sparse covariance matrices. Their main results showed

$$\inf_{\tilde{\Sigma}} \sup_{\Sigma\in\mathcal{U}_t(q,k_0,M_0)} E\|\tilde{\Sigma}-\Sigma\|^2 \asymp k_0^2(\frac{\log p}{n})^{1-q} \tag{2.6}$$

and

$$\inf_{\tilde{\Sigma}} \sup_{\Sigma\in\mathcal{U}_b(\alpha,C,M)} E\|\tilde{\Sigma}-\Sigma\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n} \tag{2.7}$$

The thresholding estimator $T_h(\hat{\Sigma})$ with threshold $h = O(\sqrt{\frac{\log p}{n}})$ achieves the optimal convergence rate for class $\mathcal{U}_t(q, k_0, M_0)$, and banded estimator $B_k(\hat{\Sigma})$ with bandwidth $k = O(n^{1/(2\alpha+1)})$ achieves optimal convergence rate for class $\mathcal{U}_b(\alpha, C, M)$. Both estimators can be made adaptable.

Results from T. Cai et al (CZZ2010-1) (CZ2012-1) showed convergence rates of approximately bandable classes are in general faster than the rates of classes with l_q constraints on rows. To do this, they showed that the banded estimator would behave as the empirical estimator for small blocks with size equivalent to the bandwidth. This suggests a natural question: suppose the population covariance matrix is not originally bandable, however under certain unknown permutation of its indices, it could be permuted into a approximately bandable structure, is it possible to obtain the same convergence rate as if the permutation were known? The answer is yes for some classes. We are particularly interested in those approximately block diagonal after permutation. Denote the class in consideration as $\mathcal{F}$. Our goal is to construct and analyze an estimator $\tilde{\Sigma}^*$ with spectral norm convergence rate minimax optimal:

$$\sup_{\Sigma\in\mathcal{F}} E\|\tilde{\Sigma}^*-\Sigma\|^2 \asymp \inf_{\tilde{\Sigma}} \sup_{\Sigma\in\mathcal{F}} E\|\tilde{\Sigma}-\Sigma\|^2 \tag{2.8}$$

The idea is to recover the unknown permutation to construct approximately block diagonal empirical covariance matrices. Then as with banding, we will show estimates keeping entries within blocks and ignoring others would achieve the optimal rates.

The problem, as we pose it, is to recover unknown blocks of variables with relatively higher correlations among themselves than outside. There are several closely related problems that have been extensively studied in the literatures. For example, sparse PCA considers the following NP hard problem

$$\hat{v} = \arg\max_{\|v\|_0\le k,\ \|v\|=1} v^t\hat{\Sigma}v \tag{2.9}$$

The goal is to estimate the first principal component (eigenvector of $\hat{\Sigma}$ corresponding to largest eigenvalue) under the assumption that the eigenvector has at most k nonzero entries. Using the Lasso by Tibshirani (T1996-1), Jolliffe et al (J2003-1) proposed SCoTLass algorithm which replaces the constraint on l_0 norm by l_1 norm. Zou, Hastie and Tibshirani (ZHT2006-1) proposed the SPCA algorithm which derives sparse principal components by solving self-constrained regression regularized by l_1 norm. d'Aspremont et al (DA2007-1) proposed the DSPCA algorithm which relaxes the l_0 constraint and transform the original problem to a semi-definite program(SDP):

$$\hat{V} = \arg \max_{V \succeq 0,\ tr(V)=1} tr(\hat{\Sigma} V) - \rho_n \sum_{i,j} |V_{ij}| \tag{2.10}$$

Amini and Wainwright (AW2009-1) analyzed the statistical properties of this method over a class of spike models

$$\mathcal{E}_\beta = \{\Sigma : \Sigma = \beta z z^t + \begin{bmatrix} I_{k\times k} & 0 \\ 0 & \Gamma_{p-k} \end{bmatrix}, \lambda_{\max}(\Gamma_{p-k}) \le 1, z_i = \pm \frac{1}{\sqrt{k}}\} \tag{2.11}$$

They showed that, under some mild assumptions, and if the solution of 2.10 is rank 1, then it agrees with the solution of the exact problem and achieves the global optima. Furthermore, if the sample size $n = ck \log p$ for sufficiently large constant c, block can be recovered with probability going to 1 for any block size $k \ge c' \log p$. However, as we have shown in simulation, the essential existence of rank 1 solution is usually not satisfied, and the SDP approach is not rate optimal. Xiaotong Yuan and Tong Zhang (YZ2011-1) proposed a truncated power method called $TPower$ and analyzed its property with more general assumptions on covariance matrices. Their method is very similar to our algorithm FB and $FBRec$ defined in Chapter 3, except that we do not re-weight the solution and we use different input matrix other than using the empirical covariance matrix directly. As the goals and analysis are quite different, our results are not directly comparable. Their goal is to recover largest sparse eigenvector and the analysis requires conditions on eigen-gap between the largest and second largest eigenvalue. We consider a different multiple block model and do not require any eigen-gap or rank conditions.

In Chapter 3, we formally setup the problem and define the appropriate class of covariance matrices, then propose corresponding estimator and algorithms. Our main results are in Chapter 4, where we analyze the statistical and computational properties of the proposed estimator and algorithms. We show that under some regularity conditions, if $k = O(\log p), n \ge O(k^2)$, and most entries within blocks have signal strength at least $O(\frac{1}{\sqrt{n}})$, which is weaker comparing to $O(\sqrt{\frac{\log p}{n}})$ signal strength required by thresholding, then our algorithm is rate optimal for support recovery, and the induced estimator achieves minimax optimal convergence rate in spectral norm. We also upper bound the worst case computational complexity of the algorithm with high probability. In Chapter 5, we show by simulations of the spike model and multiple block model to confirm that the proposed estimator is both statistically and computationally efficient.

Chapter 3

Method

3.1 Problem Setup

Recall that we are particularly interested in a class, denote as $\mathcal{F}$, of covariance matrices approximately block diagonal after unknown permutation. Our goal is to construct and analyze an estimator $\tilde{\Sigma}^*$ with spectral norm convergence rate minimax optimal:

$$\sup_{\Sigma\in\mathcal{F}} E\|\tilde{\Sigma}^* - \Sigma\|^2 \asymp \inf_{\tilde{\Sigma}} \sup_{\Sigma\in\mathcal{F}} E\|\tilde{\Sigma} - \Sigma\|^2 \tag{3.1}$$

The idea is to recover the unknown permutation to construct approximately block diagonal empirical covariance matrices. Then as with banding, we will show estimates keeping entries within blocks and ignoring others would achieve the optimal rates.

Next we define exactly the class $\mathcal{F}$ of covariance matrices approximately block diagonal after unknown permutation. Denote the set of indices as $[p] = \{1, 2, ..., p\}$. A block B is a set of pair indices if $\exists J \subset [p]$ s.t. $B = J \times J = \{(i, j) : i, j \in J\}$. Define the support of any set of pair indices B as $\mathbb{J}(B) = \{i : \exists j \text{ s.t. } (i, j) \in B \text{ or } (j, i) \in B\}$. Let $\mathbb{I}(.)$ denote the indicator function. B is an approximate block if $B \in \mathcal{B}(k, M, \epsilon)$:

$$\mathcal{B}(k, M, \epsilon) = \{B : Mk \geq |\mathbb{J}(B)| \geq k, \sum_{j\in J(B)} \mathbb{I}((i, j) \notin B) \leq \epsilon k \text{ for } \forall i \in \mathbb{J}(B)\} \tag{3.2}$$

k is the order of sizes of blocks and will grow with dimension p. M and ϵ are parameters and considered constants. ϵ is a measurement of proximity of B to a block. If ϵ is close to 0, then most $(i, j) \in \mathbb{J}(B) \times \mathbb{J}(B)$ are in B, and B is close to a block. M is a constraint so that all blocks are at the same order. A set of approximate blocks $\{B_l\}_{l=1}^m$ are approximately block diagonal if $\{B_l\}_{l=1}^m \in \mathbb{B}(k, M, \epsilon, \delta)$:

$$\mathbb{B}(k, M, \epsilon, \delta) = \{\{B_l\}_{l=1}^m : B_l \in \mathcal{B}(k, M, \epsilon), |\mathbb{J}(B_l) \cap (\cup_{i\neq l}\mathbb{J}(B_i))| \leq \delta k \tag{3.3}$$

$$\exists \text{ partition } L_1 \sqcup L_2 = [m] \text{ s.t. } B_i \cap B_j = \emptyset \text{ for } \forall (i, j) \in (L_1 \times L_1) \cup (L_2 \times L_2)\} \tag{3.4}$$

where δ is similar to ϵ and is a measurement of overlaps among blocks. If δ is close to 0, then the blocks have little overlap and are close to block diagonal. The condition

$$\exists \text{ partition } L_1 \sqcup L_2 = [m] \text{ s.t. } B_i \cap B_j = \emptyset \text{ for } \forall (i,j) \in (L_1 \times L_1) \cup (L_2 \times L_2)$$

could be relaxed to that $\exists$ constant q uniformly for all (n, p, k) in consideration,

$$\exists \text{ partition } \sqcup_{i=1}^{q} L_i = [m] \text{ s.t. } B_i \cap B_j = \emptyset \text{ for } \forall (i,j) \in \cup_{i=1}^{q} (L_i \times L_i)$$

This condition is to regularize the way blocks intersecting each other, for example, it excludes counter example proposed by El Karoui (EK2007-1) that Σ being diagonal with all other non-zero entries $\frac{1}{\sqrt{p}}$ only in the first row and the first column. We will describe how the ranges of M, ϵ and δ affect the estimator in Chapter 4. Recall that $\mathcal{S}_p^+$ is the set of $p \times p$ positive definite matrices, and Π_p is the set of all permutations of $[p]$. The class of covariance matrices we are interested in is defined as follows:

$$\mathcal{F}(\lambda, k, m, M, \epsilon, \delta) = \{\Sigma \in \mathcal{S}_p^+ : \exists \pi \in \Pi_p, \exists \{B_l\}_{l=1}^m \in \mathbb{B}(k, M, \epsilon, \delta) \tag{3.5}$$

$$\text{s.t. } \Sigma_{\pi(i)\pi(j)} = a_{ij} \mathbb{I}((i,j) \in \cup_{l=1}^m B_l) \text{ with } |a_{ij}| > \lambda \text{ for } i \neq j, \Sigma_{ii} = 1\} \tag{3.6}$$

It requires that under unknown permutation π, Σ can be permuted to approximately block diagonal with small overlaps among the blocks, and most entries within blocks are of signal strength at least λ. WLOG, we also assume $\Sigma_{ii} = 1$ for all i. Recall our goal is to construct an estimator $\tilde{\Sigma}^*$ with minimax optimal convergence rate in spectral norm:

$$\sup_{\Sigma \in \mathcal{F}(\lambda,m,k,M,\epsilon,\delta)} E\|\tilde{\Sigma}^* - \Sigma\|^2 \asymp \inf_{\tilde{\Sigma}} \sup_{\Sigma \in \mathcal{F}(\lambda,m,k,M,\epsilon,\delta)} E\|\tilde{\Sigma} - \Sigma\|^2 \tag{3.7}$$

The key parameter is the signal strength λ, since the larger λ the easier the problem. An extreme case would be $\lambda = O(\sqrt{\frac{\log p}{n}})$. In this case if $\lambda = c\sqrt{\frac{\log p}{n}}$ with a large enough constant c, the thresholding estimator $T_h(\hat{\Sigma})$ with threshold $h = O(\sqrt{\frac{\log p}{n}})$, see El Karoui (EK2007-1), Bickel and Levina (BL2008a-1)(BL2008b-1), achieves the optimal convergence rate. Our main result pushes to $\lambda = O(\frac{1}{\sqrt{n}})$ but for more specific classes with underlying block structures.

3.2 Estimator

In this section we construct a block recovery estimator with corresponding algorithms. Inspired by the analysis of banding, see T. Cai et al (CZZ2010-1), one way to estimate approximately block diagonal Σ at fastest rate would be keeping empirical estimates inside the blocks and ignore the remaining entries. Specifically, our estimator $\tilde{\Sigma}^*$ of Σ is constructed as follows:

$$\tilde{\Sigma}^*_{ij} = \hat{\Sigma}_{ij} \mathbb{I}((i,j) \in \cup_{l=1}^m \hat{B}_l) \text{ with } \hat{B}_l = \hat{J}_l \times \hat{J}_l \tag{3.8}$$

where $\hat{\Sigma}$ is the the empirical covariance matrix, and $\{\hat{J}_l\}_{l=1}^m$ are our estimate for the support of the blocks. Since permutation π is unknown, the main difficulty is to construct good estimates $\hat{J}_l$ of J_l for short of $\mathbb{J}(B_l)$. Note that the estimator is invariant under any particular labeling as long as $\{\hat{B}_l\}_{l=1}^m = \{J_l \times J_l\}_{l=1}^m$. We construct estimates for supports for $l = 1, 2, ..., m$ recursively,

$$\widehat{\mathbb{J}(B_l)} = \hat{J}_l(\theta_1, \theta_2, t) = \arg\max_J \{|J| \geq k : \frac{1}{|J|} \sum_{j \in J} \mathbb{I}(|\hat{\Sigma}_{ij}| < t) \leq \theta_1 \text{ for } \forall i \in J, |J \cap (\cup_{i=1}^{l-1} \hat{J}_i)| \leq \theta_2 k\} \tag{3.9}$$

where θ_1, θ_2, t are input parameters and their values depend on λ and $\max(\epsilon, \delta)$. The choices of θ_1, θ_2, t depend on k and λ, and the specific form is given in Theorem 4.2.1.

The main issue for support recovery is computational cost. One may recognize that it looks similar to the hidden-clique recovery problem and k-densest subgraph problem, which are NP-hard in general. Indeed, consider the graph $G(\mathcal{V}, \mathcal{E})$ induced by empirical covariance matrix with vertices $\mathcal{V} = [p]$ and edges $\mathcal{E} = \{(i, j) : |\hat{\Sigma}_{ij}| > h\}$ for some threshold h. Denote E_h the corresponding adjacency matrix:

$$E_h = (\mathbb{I}((i, j) \in \mathcal{E}))_{1 \leq i,j \leq p} = (\mathbb{I}(|\hat{\Sigma}_{ij}| > h))_{1 \leq i,j \leq p} \tag{3.10}$$

As will be shown later, then support recovery problem (3.9) for a single block is equivalent to

$$\hat{J}_1 = \arg \max_{J: |J| = |\mathbb{J}(B_1)|} \sum_{i,j \in J} (E_h)_{ij} \tag{3.11}$$

In general, if E is adjacency matrix of an arbitrary graph, this is the densest k subgraph problem and computationally intractable. However, G here is induced by thresholding empirical covariance matrix and not completely arbitrary.

3.3 Algorithm

For given threshold h, consider two matrices as potential inputs for algorithms:

$$E_h = (\mathbb{I}(|\hat{\Sigma}_{ij}| > h))_{1 \leq i,j \leq p} \tag{3.12}$$

and

$$W_h = (|\hat{\Sigma}_{ij}| \mathbb{I}(|\hat{\Sigma}_{ij} > h| \text{ and } i \neq j))_{1 \leq i,j \leq p} \tag{3.13}$$

Let A denote the generic input matrix. Our main results in Chapter 4 are based on input $A = E_h$. However, as shown in simulated experiments in Chapter 5, input $A = W_h$ is better in simulation. Specifically, algorithm $FBRec$ is proposed to recover a single block, and $FBAll$ recovers multiple blocks by repeatedly applying $FBRec$. The pseudo code is as follows:

$FB(A, J, k, t)$

$p \leftarrow$ dimension of A
for $i = 1$ to t **do**
  $R_l \leftarrow \sum_{j \in J} A_{jl}$ for $l = 1, ..., p$
  $J' \leftarrow$ indices of the top k largest elements of $\{R_l\}_{l=1}^{p}$
  if $J == J'$ **then**
    break
  else
    $J \leftarrow J'$
  end if
end for
return J

$FBS(A, J, \theta_b, \theta_r, t)$

$J' \leftarrow FB(A, J, |J| + 1, t)$
while $\min_{i \in J'} \frac{1}{|J'|} \sum_{j \in J'} A_{ij} > \theta_r$ and $\frac{1}{|J'|^2} \sum_{i,j \in J'} A_{ij} > \theta_b$ **do**
  $J \leftarrow J'$
  $J' \leftarrow FB(A, J, |J| + 1, t)$
end while
return J

$FBRec(A, L, J, k, s, \theta_b, \theta_r, t)$

if $s > 0$ **then**
  for $l = 1$ to $|L|$ **do**
    $(J_0, L', b) \leftarrow FBRec(A, \{j \in L : A_{L_l j} \neq 0\}, J \cup \{L_l\}, k, s - 1, \theta_b, \theta_r, t)$
    if $b > \theta_b$ and $\min_{i \in J_0} \frac{1}{|J_0|} \sum_{j \in J_0} A_{ij} > \theta_r$ and $|J_0| \geq k$ **then**
      BREAK
    end if
  end for
else
  $b \leftarrow 0$
  for $l = 1$ to $|L|$ **do**
    $J' \leftarrow FB(A, L_l \cup J, k, t)$
    if $\frac{1}{|J'|^2} \sum_{i,j \in J'} A_{ij} > b$ **then**
      $b \leftarrow \frac{1}{|J'|^2} \sum_{i,j \in J'} A_{ij}$
      $J_0 \leftarrow J'$
      if $b > \theta_b$ and $\min_{i \in J_0} \frac{1}{|J_0|} \sum_{j \in J_0} A_{ij} > \theta_r$ **then**
        $J_0 \leftarrow FBS(A, J_0, \theta_b, \theta_r, t)$
        break
      end if

end if
end for
end if
$L \leftarrow \{L_j\}_{j>l}$
return (J_0, L, b)

Remark: For more efficient computation, the above step
$J' \leftarrow FB(A, L_l \cup J, k, t)$
could be replaced by:
$J' \leftarrow FB(A_{L\cup J, L\cup J}, L_l \cup J, \theta_r^{|J|}k, t)$
The proof of algorithm correctness with this replacement would follow from the proof of Theorem 4.4.1 but more technical. For simplicity, Theorem 4.4.1 proves algorithm correctness without this replacement.

FBAll$(A, k, s, \theta_b, \theta_r, t)$
$p \leftarrow$ dimension of A
$L \leftarrow [p]$
$l \leftarrow 0$
while $|L| \geq k$ **do**
$(J_0, L, b) \leftarrow FBRec(A, L, \emptyset, k, s, \theta_b, \theta_r, t)$
if $b > \theta_b$ and $\min_{i\in J_0} \frac{1}{|J_0|}\sum_{j\in J_0} A_{ij} > \theta_r$ and $J_0 \notin \{B_j\}_{j=1}^{l}$ **then**
$l \leftarrow l + 1$
$B_l \leftarrow J_0$
$L \leftarrow L \backslash J_0$
end if
end while

The idea for $FBRec(A, [p], \emptyset, k, s, \theta_b, \theta_r, t)$ is to exhaustively search over all $|J| = s + 1$ satisfying

$$A_{ij} \neq 0 \text{ for } \forall i, j \in J \tag{3.14}$$

and $FB(A, J, k, t)$ is called for each such J. As shown in Theorem 4.4.1, in order for $FBRec()$ to success with high probability, $s = \eta k$ is required with $\eta > 0$ uniformly, i.e. $s = O(k)$. When $s = 0$, $FBRec()$ is very similar to $TPower$ by X.T. Yuan and T. Zhang (YZ2011-1) with special initialization for the spike model in 2.11. $FB()$ iteratively updates J by top k largest row sum R_l over J. As it does not necessarily converge, t upper bounds the number of iterations. As we will show in Theorem 4.4.1, if start with J a subset of some block J_l and $|J| = s + 1$ is sufficiently large, which always happens at some step of $FBRec()$, then

with high probability the algorithm $FB()$ converges in 1 iteration and recovers $|J_0| = k$ with J_0 a subset of J_l. Hence t could be set to small $O(1)$ constant. Recall k is the lower bound for block size. Given J_0 found by $FB()$ is a subset of a correct block, $FBS(A, J_0, \theta_r, \theta_s, t)$ recovers the full size of the block with high probability, where parameters θ_r, θ_b are chosen in Theorem 4.4.1. Finally $FBAll()$ repeatedly calls $FBRec()$ to recover all the blocks. In Theorem 4.4.1, we also provide an upper bound on worst case computational complexity of $FBAll()$. In general the algorithm takes exponential time. However, for a reasonable range of p being several thousands, the algorithm runs efficiently This is also verified by simulation in Chapter 5.

Chapter 4

Analysis

4.1 Distributional Assumptions

In this chapter we present our main results regarding convergence rates and computational complexity of proposed estimators and algorithms. Suppose $X \in \mathbb{R}^{n\times p}$ is observed, where the rows $X_1, X_2, ..., X_n \in \mathbb{R}^{1\times p}$ are i.i.d. mean 0 p-variate random variable with covariance matrix Σ. Throughout this chapter, denote $\hat{\Sigma} = \frac{X^T X}{n}$, which is the dominant component of the empirical covariance matrix. In addition to the assumptions that $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$, we make two distributional assumptions about X_i for $\forall i$:

- Assumption 1. $E[X_i] = 0$, $\exists \rho_1$ s.t. for $\forall t > 0$, $\forall v \in \mathbb{R}^p$ with $\|v\|_2 = 1$

$$P(|X_i v| > t) \leq e^{-\rho_1 t^2} \tag{4.1}$$

- Assumption 2. $\exists \rho_2, d_2 > 0$ s.t. for $\forall t < d_2$, $\forall v, w \in \mathbb{R}^p$ with $\|w\|_2 = \|v\|_2 = 1$, $\forall B \subset [p] \times [p]$,

$$P(|w^T((X_i^T X_i)_B - \Sigma_B)v| > t) < e^{-\rho_2 t^2} \tag{4.2}$$

Note Assumption 1 is the standard sub-Gaussian assumption with restriction on the largest eigenvalue of Σ being bounded above by constant. Assumption 2 is similar to sub-exponential assumption but slightly stronger. A fact is that if X follows a Gaussian distribution, then Assumption 2 is implied by Assumption 1. As a special case, the only condition for Gaussian distribution would be the largest eigenvalue of Σ being bounded above by constant. These distributional assumptions are made due to development of technical convergence rates. Similar methods and techniques could be applied to other classes of distributions to obtain different convergence rates.

4.2 Optimal Support Recovery

Theorem 4.2.1. *Suppose i.i.d mean 0 p-variate sub-Gaussian $X_1, ..., X_n$ satisfying Assumption 1 and Assumption 2, with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$.*

- *Upper Bound: If uniformly for all (k, n, p), $k \geq \log p, \lambda = \frac{C}{\sqrt{n}}, \frac{n}{k^2} > \frac{1+\log 2}{\rho_2 d_2^2}$, and*

$$\max(\epsilon, \delta) + \frac{8(2+\log 2)}{\rho C^2} + \frac{4\sqrt{1+\log 2}}{C\sqrt{\rho_2}} + \frac{(M+1)^2(\max(\epsilon,\delta) + \frac{8(2+\log 2)}{\rho C^2})^2}{(M+1)(\max(\epsilon,\delta) + \frac{8(2+\log 2)}{\rho C^2}) - \delta} < 1 \tag{4.3}$$

  where ρ from Lemma 4.2.2 depends on ρ_1. Then $\exists \gamma_1, \gamma_2, \alpha$ s.t.

$$P(\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m = \{J_l\}_{l=1}^m) < e^{-[\rho C^2(\gamma_1 - \max(\delta,\epsilon))/8 - \log 2 - 2]k} + e^{-[\rho_2 C^2(\gamma_2 - \frac{\alpha^2}{\alpha-\delta})^2/16 - \log 2 - 1]k^2} \tag{4.4}$$

$$\to 1 \text{ as } (k, n, p) \to \infty \tag{4.5}$$

- *Lower Bound: On the other hand, if*

$$\frac{n}{\log p} < \frac{1 + k\lambda}{k\lambda^2} \tag{4.6}$$

  then the probability of error of any method is at least $\frac{1}{2}$.

- *Optimal case: If $k = O(\log p)$, $\lambda = O(\frac{1}{\sqrt{n}})$, $\frac{n}{k^2} \geq O(1)$, then $\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m$ is rate optimal, i.e. if $k = c_1 \log p$ with $c_1 \geq 1, \lambda = \frac{c_2}{\sqrt{n}}, \frac{n}{k^2} \geq c_3$, then $\exists c_4 < c_5$ s.t.*

$$c_2 > c_5 \Rightarrow P(\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m = \{J_l\}_{l=1}^m) \to 1 \text{ as } (k, n, p) \to \infty \tag{4.7}$$

$$c_2 < c_4 \Rightarrow P(\text{error of any method}) > \frac{1}{2} \tag{4.8}$$

Let us introduce some lemma before the proof of Theorem 4.2.1.

Lemma 4.2.2. *Let $e_i = (0, ..., 0, 1, 0, ..., 0)^T \in \mathbb{R}^p$ with 1 in the i-th entry. Suppose X satisfies Assumption 1. Suppose $E[X_i^T X_i] = \Sigma$, $\hat{\Sigma} = \frac{X^T X}{n}$. Then $\exists$ ρ, d uniform for all n, p, v with $\|v\|_2 = 1$, and for $\forall$ $i \in [p]$, $\forall J \subset [p]$, $\forall t < d$*

$$P(|e_i^T(\hat{\Sigma}_{iJ} - \Sigma_{iJ})v| > t) < e^{-\rho n t^2} \tag{4.9}$$

Proof. $e_i^T\hat{\Sigma}_{iJ}v = e_i^T\frac{(X^TX)_{iJ}}{n}v = \frac{1}{n}X_i(\sum_{j\in J}v_jX_j)$. Since X_i and $\sum_{j\in J}v_jX_j$ are sub-Gaussian with $O(1)$ variance by assumption, and

$$Var(e_i^T(X_{(1)}^TX_{(1)})_{iJ}v) \le E[X_{1i}^4] + E[(\sum_{j\in J}v_jX_{1j})^4] = O(1) \tag{4.10}$$

This implies $e_i^T\hat{\Sigma}_{iJ}v$ is sub-exponential with variance $O(\frac{1}{n})$, which completes the proof.

□

Lemma 4.2.3. *Same assumptions in Lemma 4.2.2, for $\forall i \in [p], \forall J \subset [p]$, denote*

$$\mathcal{C}_i(t, J) = \frac{1}{|J|}\sum_{j\in J}\mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t) \tag{4.11}$$

then for $\forall\gamma \in (0, 1)$ and $\forall t < \frac{d}{\sqrt{\gamma|J|/2}}$,

$$P(\mathcal{C}_i(t, J) > \gamma) < 2^{|J|}e^{-\rho n\gamma|J|t^2/2} \tag{4.12}$$

Proof. $\mathcal{C}_i(t, J) > \gamma$ implies that $\exists L \subseteq J$ with $|L| = \frac{1}{2}\gamma|J|$ s.t.

$$|\sum_{j\in L}(\hat{\Sigma}_{ij} - \Sigma_{ij})| > |L|t \tag{4.13}$$

By union bound, and apply Lemma 4.2.2 to $v = (\frac{1}{\sqrt{|L|}})_{j\in L}$, i.e. $v_j = \frac{1}{\sqrt{|L|}}$ for $j \in L$ and $v_j = 0$ for $j \notin L$, we have for $\forall t < \frac{d}{\sqrt{|L|}} = \frac{d}{\sqrt{\gamma|J|/2}}$,

$$P(\mathcal{C}_i(t, J) > \gamma) \le P(\cup_{L:|L|=\gamma|J|/2}\{\frac{1}{\sqrt{|L|}}|\sum_{j\in L}(\hat{\Sigma}_{ij} - \Sigma_{ij})| > \frac{|L|t}{\sqrt{|L|}}\}) \tag{4.14}$$

$$\le \binom{|J|}{|L|}P(\frac{1}{\sqrt{|L|}}|\sum_{j\in L}(\hat{\Sigma}_{ij} - \Sigma_{ij})| > \sqrt{|L|}t) \tag{4.15}$$

$$\le 2^{|J|}e^{-\rho n|L|t^2} \tag{4.16}$$

$$\le 2^{|J|}e^{-\rho n\gamma|J|t^2/2} \tag{4.17}$$

□

Lemma 4.2.4. *Suppose X satisfies Assumption 1 with covariance matrix Σ,*

$$\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$$

Recall that $\{\mathbb{J}(B_l)\}_{l=1}^m$ are support of the blocks, denote $J_l = \mathbb{J}(B_l)$ for short. For $t \in (0, \lambda)$, let

$$H_1(t, \gamma) = \{\exists J_l, \exists i \notin J_l \text{ s.t. } \frac{1}{|J_l|}\sum_{j\in J_l}\mathbb{I}(|\hat{\Sigma}_{ij}| > t) > \gamma\} \tag{4.18}$$

$$H_2(t,\gamma) = \{\exists J_l, \exists i \in J_l \ s.t. \ \frac{1}{|J_l|}\sum_{j\in J_l} \mathbb{I}(|\hat{\Sigma}_{ij}| < t) > \gamma\} \tag{4.19}$$

if $\gamma > \max(\delta,\epsilon)$, $t \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma-\epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma-\delta)k/2}}))$,

$$P(H_1(t,\gamma) \cup H_2(t,\gamma)) \leq \sum_{l=1}^{m} p 2^{|J_l|} e^{-\rho n(\gamma|J_l| - \max(\delta,\epsilon)k)\min(t^2,(\lambda-t)^2)/2} \tag{4.20}$$

if $\rho n\gamma \min(t^2,(\lambda-t)^2)/2 > \log 2$,

$$P(H_1(t,\gamma) \cup H_2(t,\gamma)) < mpe^{-(\rho n(\gamma - \max(\delta,\epsilon))\min(t^2,(\lambda-t)^2)/2 - \log 2)k} \tag{4.21}$$

Proof. $\Sigma \in \mathcal{F}(\lambda,k,m,M,\epsilon,\delta)$ implies

$$\sum_{j\in J_l} \mathbb{I}(|\Sigma_{ij}| \neq 0) \leq \delta k \text{ for } \forall i \notin J_l \tag{4.22}$$

$$\sum_{j\in J_l} \mathbb{I}(|\Sigma_{ij}| = 0) \leq \epsilon k \text{ for } \forall i \in J_l \tag{4.23}$$

H_1 implies $\exists J_l$, $\exists i \notin J_l$ s.t.

$$\mathcal{C}_i(t, J_l) = \frac{1}{|J_l|}\sum_{j\in J_l} \mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t) > \gamma - \frac{\delta k}{|J_l|} \tag{4.24}$$

Lemma 4.2.3 implies for $\forall t < \frac{d}{\sqrt{(\gamma|J_l| - \delta k)/2}} \leq \frac{d}{\sqrt{(\gamma-\delta)k/2}}$

$$P(H_1(t,\gamma)) \leq \sum_{l=1}^{m}\sum_{i\notin J_l} P(\mathcal{C}_i(t,J_l) > \gamma - \frac{\delta k}{|J_l|}) \tag{4.25}$$

$$\leq \sum_{l=1}^{m}(p - |J_l|)2^{|J_l|}e^{-\rho n(\gamma|J_l| - \delta k)t^2/2} \tag{4.26}$$

Similarly, H_2 implies $\exists J_l$, $\exists i \in J_l$ s.t.

$$\mathcal{C}_i(\lambda - t, J_l) = \frac{1}{|J_l|}\sum_{j\in J_l} \mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > \lambda - t) > \gamma - \frac{\epsilon k}{|J_l|} \tag{4.27}$$

Lemma 4.2.3 implies that if $\lambda - t < \frac{d}{\sqrt{(\gamma|J_l| - \epsilon k)/2}} \leq \frac{d}{\sqrt{(\gamma-\epsilon)k/2}}$,

$$P(H_2(t,\gamma)) \leq \sum_{l=1}^{m}\sum_{i\in J_l} P(\mathcal{C}_i(\lambda - t, J_l) > \gamma - \frac{\epsilon k}{|J_l|}) \tag{4.28}$$

$$\leq \sum_{l=1}^{m}|J_l|2^{|J_l|}e^{-\rho n(\gamma|J_l| - \epsilon k)(\lambda-t)^2/2} \tag{4.29}$$

Combine the results, if $\gamma > \max(\delta, \epsilon)$ and $t \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma-\epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma-\delta)k/2}}))$,

$$P(H_1(t,\gamma) \cup H_2(t,\gamma)) \leq \sum_{l=1}^{m} p 2^{|J_l|} e^{-\rho n(\gamma|J_l| - \max(\delta,\epsilon)k)\min(t^2,(\lambda-t)^2)/2} \tag{4.30}$$

□

Lemma 4.2.5. *Suppose X satisfies Assumption 1 and Assumption 2 with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$. For $\forall \alpha \in (\delta, 1)$, let*

$$\mathcal{J}_\alpha = \{J \subset [p] : k \leq |J|, \ \frac{|J \cap J_l|}{|J|} < \alpha \text{ for } l = 1, 2, ..., m\} \tag{4.31}$$

Let

$$g_\gamma(t, J) = \{\frac{1}{|J|^2} \sum_{(i,j) \in J \times J} \mathbb{I}(|\hat{\Sigma}_{ij}| > t) > \gamma\} \tag{4.32}$$

Let $G_\alpha(t,\gamma) = \cup_{J \in \mathcal{J}_\alpha} g_\gamma(t, J)$, for $\forall t < \frac{2d_2}{(\gamma - \frac{\alpha^2}{\alpha-\delta})k}$,

$$P(G_\alpha(t,\gamma)) \leq \sum_{|J|=k}^{p} p^{|J|} 2^{|J|^2} e^{-\rho_2 n(\gamma - \frac{\alpha^2}{\alpha-\delta})^2 |J|^2 t^2/4} \tag{4.33}$$

if $\rho_2 n(\gamma - \frac{\alpha^2}{\alpha-\delta})^2 t^2/4 > 1 + \log 2$ and $k \geq \log p$,

$$P(G_\alpha(t,\gamma)) \leq e^{-(\rho_2 n(\gamma - \frac{\alpha^2}{\alpha-\delta})^2 t^2/4 - \log 2 - 1)k^2} \tag{4.34}$$

Proof. For $\forall J \in \mathcal{J}_\alpha$, there are at most $\frac{|J|}{\alpha|J| - \delta k}$ many J_l having non-empty intersection with J, and the cardinalities of these non-empty intersections are upper bounded by $\alpha^2|J|^2$, which implies

$$\sum_{(i,j) \in J \times J} \mathbb{I}(\Sigma_{ij} \neq 0) \leq \frac{|J|}{\alpha|J| - \delta k} \alpha^2 |J|^2 \leq \frac{1}{\alpha - \delta} \alpha^2 |J|^2 \tag{4.35}$$

$G_\alpha(t,\gamma)$ implies $\exists J \in \mathcal{J}_\alpha$ s.t. $g_\gamma(t, J)$ happens, which implies there exist at least $\gamma|J|^2 - \frac{\alpha^2}{\alpha-\delta}|J|^2$ many $(i,j) \in J \times J$ s.t. $\Sigma_{ij} = 0$ and $|\hat{\Sigma}_{ij}| > t$. Thus $\exists L \subseteq J \times J$ with

$$|L| = \tfrac{1}{2}(\gamma - \tfrac{\alpha^2}{\alpha-\delta})|J|^2$$

s.t.

$$|\sum_{(i,j) \in L} (\hat{\Sigma}_{ij} - \Sigma_{ij})| > t|L| \tag{4.36}$$

Apply Assumption 2 to $v = (\frac{1}{\sqrt{|J|}})_{j\in J}^T$, then for $\forall t < \frac{d_2|J|}{|L|} = \frac{2d_2}{(\gamma - \frac{\alpha^2}{\alpha-\delta})|J|} \leq \frac{2d_2}{(\gamma-\frac{\alpha^2}{\alpha-\delta})k}$,

$$P(g_\gamma(t,J)) \leq \binom{|J|^2}{|L|} P(\frac{1}{|J|}|\sum_{(i,j)\in L}(\hat{\Sigma}_{ij} - \Sigma_{ij})| > \frac{t|L|}{|J|}) \tag{4.37}$$

$$= 2^{|J|^2} P(|v^T(\hat{\Sigma}_L - \Sigma_L)v| > \frac{t|L|}{|J|}) \tag{4.38}$$

$$\leq 2^{|J|^2} e^{-\rho_2 n(\frac{t|L|}{|J|})^2} \tag{4.39}$$

$$= 2^{|J|^2} e^{-\rho_2 n(\gamma - \frac{\alpha^2}{\alpha-\delta})^2|J|^2t^2/4} \tag{4.40}$$

By union bound

$$P(G_\alpha(t,\gamma)) \leq \sum_{J\in\mathcal{J}_\alpha} P(g_\gamma(t,J)) \tag{4.41}$$

$$\leq \sum_{|J|=k}^{p} \binom{p}{|J|} 2^{|J|^2} e^{-\rho_2 n(\gamma-\frac{\alpha^2}{\alpha-\delta})^2|J|^2t^2/4} \tag{4.42}$$

$$\leq \sum_{|J|=k}^{p} p^{|J|} 2^{|J|^2} e^{-\rho_2 n(\gamma-\frac{\alpha^2}{\alpha-\delta})^2|J|^2t^2/4} \tag{4.43}$$

$$\tag{4.44}$$

□

Lemma 4.2.6. *Suppose X satisfies Assumption 1 and Assumption 2 with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$. Recall that the estimator $\{\hat{J}_l\}_{l=1}^m$ for $\{\mathbb{J}(B_l)\}_{l=1}^m$ ($\{J_l\}_{l=1}^m$ for short) are constructed for $l = 1, ..., m$ recursively:*

$$\hat{J}_l(\theta_1,\theta_2,t) = \arg\max_J\{|J| \geq k : \frac{1}{|J|}\sum_{j\in J}\mathbb{I}(|\hat{\Sigma}_{ij}| < t) \leq \theta_1 \text{ for } \forall i \in J, |J \cap (\cup_{i=1}^{l-1}\hat{J}_i)| \leq \theta_2 k\} \tag{4.45}$$

Let $\theta_1 = \gamma$, $\theta_2 = \delta$, $\alpha \in (\delta, 1)$, $\gamma_1 < \frac{\alpha}{M+1}$ and $\gamma_2 < 1 - \gamma_1$, then

$$P(\{\hat{J}_l(\gamma_1,\delta,t)\}_{l=1}^m = \{J_l\}_{l=1}^m) > 1 - P(H_1(t,\gamma_1) \cup H_2(t,\gamma_1) \cup G_\alpha(t,\gamma_2)) \tag{4.46}$$

Proof. Suppose $H_1(t,\gamma_1)^c \cap H_2(t,\gamma_1)^c \cap G_\alpha(t,\gamma_2)^c$ happens. Recall

$$H_1(t,\gamma_1) = \{\exists J_l, \exists i \notin J_l \text{ s.t. } \frac{1}{|J_l|}\sum_{j\in J_l}\mathbb{I}(|\hat{\Sigma}_{ij}| > t) > \gamma_1\} \tag{4.47}$$

$$H_2(t,\gamma_1) = \{\exists J_l, \exists i \in J_l \text{ s.t. } \frac{1}{|J_l|}\sum_{j\in J_l}\mathbb{I}(|\hat{\Sigma}_{ij}| < t) > \gamma_1\} \tag{4.48}$$

$$G_\alpha(t,\gamma_2) = \{\exists J \in \mathcal{J}_\alpha \ s.t. \frac{1}{|J|^2}\sum_{i,j\in J} \mathbb{I}(|\hat{\Sigma}_{ij}| > t) > \gamma_2\} \tag{4.49}$$

$\theta_1 = \gamma_1$, $\theta_2 = \delta$ and $H_2(t,\gamma_1)^c$ imply that for $l = 1, 2, ..., m$ and for $\forall i \in J_l$, $\frac{1}{|J_l|}\sum_{j\in J_l} \mathbb{I}(|\hat{\Sigma}_{ij}| < t) \leq \theta_1$, and $|J_l \cap (\cup_{s\neq l} J_s)| \leq \theta_2 k$. This implies that $\exists\{\hat{J}_l\}_{l=1}^m$ satisfying equation 4.45 s.t. $J_l \subset \hat{J}_l$ for $l = 1, 2, ..., m$. Now it remains to show that $\{J_l\}_{l=1}^m$ is indeed the unique solution of 4.45:

Suppose $\exists\{\hat{J}_l\}_{l=1}^m$ satisfying 4.45 different from $\{J_l\}_{l=1}^m$, then $\exists \hat{J} \in \{\hat{J}_l\}_{l=1}^m$ s.t. $\hat{J} \neq J_l$ for $\forall l$, thus one of following two cases must hold:

Case 1: $\exists l$ s.t. $|J_l \cap \hat{J}| \geq \alpha|\hat{J}|$, which contradicts $H_1(t,\gamma_1)^c$ if

$$\alpha|\hat{J}| - \gamma_1|J_l| > \theta_1|\hat{J}| \tag{4.50}$$

Case 2: For $\forall l$, $|J_l \cap \hat{J}| < \alpha|\hat{J}|$, which contradicts $G_\alpha(t,\gamma_2)^c$ if

$$\gamma_2|\hat{J}|^2 < (1-\theta_1)|\hat{J}|^2 \tag{4.51}$$

If $\theta_1 = \gamma$, $\theta_2 = \delta$, then conditions in Case 1 and Case 2 can be deduced to

$$\gamma_1 < \frac{\alpha}{M+1} \text{ and } \gamma_2 < 1 - \gamma_1 \tag{4.52}$$

Hence if condition 4.52 is satisfied, the event that $\exists\{\hat{J}_l\}_{l=1}^m$ different from $\{J_l\}_{l=1}^m$ implies contradiction, i.e. $H_1(t,\gamma_1)^c \cap H_2(t,\gamma_1)^c \cap G_\alpha(t,\gamma_2)^c$ implies that $\{\hat{J}_l(\gamma_1,\delta,t)\}_{l=1}^m = \{J_l\}_{l=1}^m$ is the unique solution of 4.45. □

Lemma 4.2.7. *(Amini and Wainwright (AW2009-1)) Consider the spike model $\mathcal{E}_\beta$ defined in 2.11. If*

$$\frac{n}{k\log(p-k)} < \frac{1+\beta}{\beta^2} \tag{4.53}$$

Then the probability of error of any method is at least $\frac{1}{2}$.

Proof. Refer to Amini and Wainwright (AW2009-1) Theorem 3. □

Proof of Theorem 4.2.1:

Proof. Combine Lemma 4.2.4, Lemma 4.2.5, Lemma 4.2.6, if $\alpha \in (\delta, 1)$, $k \geq \log p$, and

$$\gamma_1 < \frac{\alpha}{M+1} \tag{4.54}$$

$$\gamma_2 < 1 - \gamma_1 \tag{4.55}$$

$$t \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma_1 - \epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma_1 - \delta)k/2}})) \tag{4.56}$$

$$t < \frac{2d_2}{(\gamma_2 - \frac{\alpha^2}{\alpha - \delta})k} \tag{4.57}$$

$$2 + \log 2 < \rho n(\gamma_1 - \max(\delta, \epsilon)) \min(t^2, (\lambda - t)^2)/2 \tag{4.58}$$

$$1 + \log 2 < \rho_2 n(\gamma_2 - \frac{\alpha^2}{\alpha - \delta})^2 t^2/4 \tag{4.59}$$

then

$$P(\{\hat{J}_l(\gamma_1, \delta, t)\}_{l=1}^m \neq \{J_l\}_{l=1}^m) \leq P(H_1(t, \gamma_1) \cup H_2(t, \gamma_1)) + P(G_\alpha(t, \gamma_2)) \tag{4.60}$$

$$\leq mpe^{-[\rho n(\gamma_1 - \max(\delta,\epsilon)) \min(t^2, (\lambda - t)^2)/2 - \log 2]k} \tag{4.61}$$

$$+ e^{-[\rho_2 n(\gamma_2 - \frac{\alpha^2}{\alpha - \delta})^2 t^2/4 - \log 2 - 1]k^2} \tag{4.62}$$

$$\leq e^{-[\rho n(\gamma_1 - \max(\delta,\epsilon)) \min(t^2, (\lambda - t)^2)/2 - \log 2 - 2]k} \tag{4.63}$$

$$+ e^{-[\rho_2 n(\gamma_2 - \frac{\alpha^2}{\alpha - \delta})^2 t^2/4 - \log 2 - 1]k^2} \tag{4.64}$$

If $t = \frac{\lambda}{2} = \frac{C}{2\sqrt{n}}$, and $\frac{n}{k^2} > C_0$ uniformly for all (n, k), then condition 4.57 is asymptotically stronger than 4.56, i.e. 4.57 implies 4.56 for (n, k) big enough. Now conditions 4.54 – 4.59 can be simplified as

$$1 > \gamma_1 + \gamma_2 \tag{4.65}$$

$$\frac{\alpha}{M+1} > \gamma_1 > \max(\epsilon, \delta) + \frac{8(2 + \log 2)}{\rho C^2} \tag{4.66}$$

$$\frac{\alpha^2}{\alpha - \delta} + \frac{4d_2\sqrt{n}}{Ck} > \gamma_2 > \frac{\alpha^2}{\alpha - \delta} + \frac{4\sqrt{1 + \log 2}}{C\sqrt{\rho_2}} \tag{4.67}$$

Hence $\exists \gamma_1, \gamma_2$ satisfying 4.65 – 4.67 if $\exists \alpha \in (\delta, 1)$ s.t.

$$1 > \max(\epsilon, \delta) + \frac{8(2 + \log 2)}{\rho C^2} + \frac{4\sqrt{1 + \log 2}}{C\sqrt{\rho_2}} + \frac{\alpha^2}{\alpha - \delta} = A + \frac{\alpha^2}{\alpha - \delta} \tag{4.68}$$

$$\alpha > (M+1)(\max(\epsilon, \delta) + \frac{8(2 + \log 2)}{\rho C^2}) = B \tag{4.69}$$

$$\frac{4d_2\sqrt{n}}{Ck} > \frac{4\sqrt{1 + \log 2}}{C\sqrt{\rho_2}} \tag{4.70}$$

Note that $M \geq 1$ and $B \geq 2\delta$. Since $\frac{d}{d\alpha}(\frac{\alpha^2}{\alpha-\delta}) = \frac{(\alpha-2\delta)\alpha}{(\alpha-\delta)^2} > 0$ for $\forall \alpha > 2\delta$, hence if $A + \frac{B^2}{B-\delta} < 1$, then $\exists \alpha \in (\delta, 1)$ satisfying 4.68 and 4.69.

Thus if uniformly for all (k, n, p), $k \geq \log p, \lambda = \frac{C}{\sqrt{n}}, \frac{n}{k^2} > C_0$, and if

$$1 > A + \frac{B^2}{B-\delta} = \max(\epsilon, \delta) + \frac{8(2+\log 2)}{\rho C^2} + \frac{4\sqrt{1+\log 2}}{C\sqrt{\rho_2}} \tag{4.71}$$

$$+ \frac{(M+1)^2(\max(\epsilon, \delta) + \frac{8(2+\log 2)}{\rho C^2})^2}{(M+1)(\max(\epsilon, \delta) + \frac{8(2+\log 2)}{\rho C^2}) - \delta} \tag{4.72}$$

$$\frac{d_2^2 n}{k^2} > \frac{1+\log 2}{\rho_2} \tag{4.73}$$

then $\exists \gamma_1, \gamma_2, \alpha$ s.t. as $(k, n, p) \to \infty$,

$$P(\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m = \{J_l\}_{l=1}^m) \to 1 \tag{4.74}$$

This completes the proof of Upper Bound in Theorem 4.2.1.

Note that $\mathcal{E}_\beta$ with $\Gamma_{p-k} = I_{p-k \times p-k}$ is a subset of $\mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$ with $\lambda = \frac{\beta}{k}, m = 1, M = 1, \epsilon = 0, \delta = 0$. Hence Lemma 4.2.7 with $\lambda = \frac{\beta}{k}$ implies the Lower Bound in Theorem 4.2.1.

Lemma 4.2.7 also suggests that $\frac{n}{k \log p} > \frac{1+k\lambda}{k^2\lambda^2}$ must hold in order to get perfect support recovery. If $\frac{1+k\lambda}{k^2\lambda^2}$ converges, one of following three cases must hold:

- $\frac{1+k\lambda}{k^2\lambda^2} \to \infty$:

  This implies $k\lambda \to 0$. Thus $\frac{n}{k \log p} > \frac{1+k\lambda}{k^2\lambda^2}$ if and only if $\lambda^2 = \frac{C^2}{n} > \frac{\log p}{kn}$. $\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m$ is rate optimal if $k = O(\log p)$ and $\frac{n}{k^2} \to \infty$.

- $\frac{1+k\lambda}{k^2\lambda^2} \to O(1)$:

  This implies $k\lambda \to O(1)$. It is rate optimal if $k = O(\log p)$ and $n = O(k^2)$.

- $\frac{1+k\lambda}{k^2\lambda^2} \to 0$:

  This implies $k\lambda \to \infty$, contradicting $\frac{n}{k^2} > \frac{1+\log 2}{\rho_2 d_2^2}$. Hence no optimal rate is achieved.

□

4.3 Minimax Optimal Covariance Estimation

Theorem 4.2.1 shows that the support recovery estimator $\{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m$ is rate optimal for the specified range of parameters. After the block structure is recovered, we can estimate Σ with $\tilde{\Sigma}^*$ induced by $\{\hat{J}_l\}_{l=1}^m = \{\hat{J}_l(\gamma_1, \delta, \frac{\lambda}{2})\}_{l=1}^m$:

$$\tilde{\Sigma}_{ij}^* = \hat{\Sigma}_{ij} \mathbb{I}((i,j) \in \cup_{l=1}^m \hat{J}_l \times \hat{J}_l) \tag{4.75}$$

Next theorem shows $\tilde{\Sigma}^*$ is minimax optimal in spectral norm for class $\mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$ with parameters in the specified range.

Theorem 4.3.1. *Suppose i.i.d mean 0 p-variate sub-Gaussian $X_1, ..., X_n$ satisfy Assumption 1 and Assumption 2, with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$.*

If $k = O(\log p), \lambda = \frac{C}{\sqrt{n}}, \frac{n}{k^2} > \frac{1+\log 2}{\rho_2 d_2^2}$, and

$$\max(\epsilon, \delta) + \frac{8(2+\log 2)}{\rho C^2} + \frac{4\sqrt{1+\log 2}}{C\sqrt{\rho_2}} + \frac{(M+1)^2(\max(\epsilon,\delta) + \frac{8(2+\log 2)}{\rho C^2})^2}{(M+1)(\max(\epsilon,\delta) + \frac{8(2+\log 2)}{\rho C^2}) - \delta} < 1 \quad (4.76)$$

Then

$$\inf_{\tilde{\Sigma}} \sup_{\Sigma \in \mathcal{F}(\lambda,k,m,M,\epsilon,\delta)} E\| \tilde{\Sigma} - \Sigma\|^2 \asymp \frac{k}{n} \quad (4.77)$$

Proof. The proof of Theorem 4.3.1 consists of two parts: Lemma 4.3.3 shows the upper bound, and Lemma 4.3.5 shows the lower bound. □

Following are lemmas relevant to the proof of Theorem 4.3.1.

Lemma 4.3.2. *(T. Cai et al (CZZ2010-1)) For any k dimensional sub-Gaussian r.v. with covariance matrix Σ_k with largest eigenvalue upper bounded uniformly for all k, $\exists \rho_3$ s.t. for $\forall t < \rho_3$,*

$$P(\|\hat{\Sigma}_k - \Sigma_k\| > t) < 5^k e^{-\rho_3 n t^2} \quad (4.78)$$

Proof. Refer to T. Cai et al (CZZ2010-1) Lemma 3. □

Lemma 4.3.3. *Suppose all assumptions in Theorem 4.3.1 hold. Recall $\tilde{\Sigma}^*$ is induced by $\{\hat{J}_l\}_{l=1}^m$:*

$$\tilde{\Sigma}^*_{ij} = \hat{\Sigma}_{ij} \mathbb{I}((i,j) \in \cup_{l=1}^m \hat{J}_l \times \hat{J}_l) \quad (4.79)$$

Then $\exists C_0, C_1 > 0$ s.t. if $k > C_0 \log p$, then

$$\sup_{\Sigma \in \mathcal{F}(\lambda,k,m,M,\epsilon,\delta)} E[\|\tilde{\Sigma}^* - \Sigma\|^2] \leq C_1 \frac{Mk}{n} \quad (4.80)$$

Proof. Denote $A = \{\{\hat{J}_l\}_{l=1}^m = \{J_l\}_{l=1}^m\}$,

$$E[\|\tilde{\Sigma}^* - \Sigma\|^2] = E[\|\tilde{\Sigma}^* - \Sigma\|^2(\mathbb{I}_A + \mathbb{I}_{A^c})] \quad (4.81)$$

Recall that by construction of the blocks, $\exists$ partition $L_1 \sqcup L_2 = [m]$ s.t. $J_i \cap J_j = \emptyset$ for $\forall (i,j) \in (L_1 \times L_1) \cup (L_2 \times L_2)$, this implies

$$\|\tilde{\Sigma}^* - \Sigma\|\mathbb{I}_A \leq \|\sum_{l \in L_1} \hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| + \|\sum_{l \in L_2} \hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| \quad (4.82)$$

$$\leq \max_{l \in L_1} \|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| + \max_{l \in L_2} \|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| \quad (4.83)$$

$$\leq 2 \max_{l=1,...,m} \|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| \quad (4.84)$$

Denote $N^{(m)} = \max_{l=1,\dots,m} \|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\|$. Let $B = \{N^{(m)} > t\}$. Lemma 4.3.2 implies

$$P(B) = P(\max_{l=1,\dots,m} \|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| > t) \tag{4.85}$$

$$\leq m \max_{l=1,\dots,m} P(\|\hat{\Sigma}_{J_l \times J_l} - \Sigma_{J_l \times J_l}\| > t) \tag{4.86}$$

$$< m 5^{Mk} e^{-\rho_3 n t^2} \tag{4.87}$$

Let $t = c_1 \sqrt{\frac{Mk}{n}}$ with $\frac{c_1^2}{2} > \frac{\log 5}{\rho_3}$, then $\exists c_2 > 0$ s.t.

$$E[\|\tilde{\Sigma}^* - \Sigma\|^2 \mathbb{I}_A] \leq 4E[(N^{(m)})^2 (\mathbb{I}_B + \mathbb{I}_{B^c})] \tag{4.88}$$

$$\leq 4E[t^2 + (N^{(m)})^2 \mathbb{I}_B] \tag{4.89}$$

$$\leq 4(t^2 + \sqrt{E[(N^{(m)})^4] P(B)}) \tag{4.90}$$

$$\leq 4(t^2 + (Mk)^2 m 5^{Mk} e^{-\rho_3 n t^2/2}) \tag{4.91}$$

$$\leq c_2 \frac{Mk}{n} \tag{4.92}$$

Recall that Theorem 4.2.1 implies if assumptions in Theorem 4.3.1 hold, then $\exists c_3 > 0$ s.t. $P(A) < e^{-c_3 k}$. Thus if $C_0 > \frac{4}{c_3}$ and $k \geq C_0 \log p$,

$$E[\|\tilde{\Sigma}^* - \Sigma\|^2 \mathbb{I}_{A^c}] \leq \sqrt{E[\|\tilde{\Sigma}^* - \Sigma\|^4] P(A^c)} \tag{4.93}$$

$$\leq p^4 e^{-c_3 k} \tag{4.94}$$

$$\leq p^{-(C_0 c_3 - 4)} \tag{4.95}$$

Combine the results, $\exists C_0, C_1$ s.t. if $k > C_0 \log p$, then

$$E[\|\tilde{\Sigma}^* - \Sigma\|^2] = E[\|\tilde{\Sigma}^* - \Sigma\|^2 (\mathbb{I}_A + \mathbb{I}_{A^c})] \tag{4.96}$$

$$\leq c_2 \frac{Mk}{n} + p^{-(C_0 c_3 - 4)} \tag{4.97}$$

$$\leq C_1 \frac{Mk}{n} \tag{4.98}$$

□

Lemma 4.3.4. *(Generalized Fano's Lemma, B. Yu (Y1997-1)) Let $\mathcal{M}_r \subset \mathcal{P}$ contains r probability measures such that for all $i \neq j$ with $i, j \leq r$*

$$d(\theta(P_i), \theta(P_j)) \geq \alpha_r \tag{4.99}$$

and

$$D(P_i \| P_j) \leq \beta_r \tag{4.100}$$

where $D(P_i\|P_j)$ denotes the Kullback-Leibler(K-L) divergence of P_j from P_i,

$$D(P_i\|P_j) = \int \ln(\frac{dP_i}{dP_j})dP_i \tag{4.101}$$

Then

$$\max_i E_{P_i}[d(\hat{\theta}, \theta(P_i))] \geq \frac{\alpha_r}{2}(1 - \frac{\beta_r + \log 2}{\log r}) \tag{4.102}$$

Proof. Refer to Assouad, Fano, and Le Cam by Bin Yu (Y1997-1). □

Lemma 4.3.5. *If $\lambda = \frac{C}{\sqrt{n}}$ with $C < \sqrt{\frac{\log p}{2k}}$, then*

$$\inf_{\tilde{\Sigma}} \sup_{\Sigma \in \mathcal{F}(\lambda,k,m,M,\epsilon,\delta)} E[\|\tilde{\Sigma} - \Sigma\|^2] \geq \frac{C^2 Mk}{16n} \tag{4.103}$$

Proof. Let $r = \binom{p}{k}$, denote the set of all subsets of $\{1, ..., p\}$ with k elements by

$$S = \{L_i \subset \{1, ..., p\} : |L_i| = k \text{ for } i = 1, ..., r\} \tag{4.104}$$

Denote $B_i \in \mathbb{R}^{p\times p}$ a block matrix with block index L_i,

$$(B_i)_{jl} = \mathbb{I}(j, l \in L_i, j \neq l) \tag{4.105}$$

and $D_i \in \mathbb{R}^{p\times p}$ a diagonal block matrix,

$$(D_i)_{jl} = \mathbb{I}(j, l \in L_i, j = l) \tag{4.106}$$

Consider P_i the join distribution of n *i.i.d.* $p-$dimensional Gaussian with covariance matrix $I_{p\times p} + \lambda B_i$

$$\mathcal{M}_r = \{P_i : \Sigma(P_i) = I_{p\times p} + \lambda B_i \text{ for } i = 1, ..., r\} \tag{4.107}$$

Denote $\Sigma_i = \Sigma(P_i)$, consider $d = \|.\|$ the operator norm, then for $\forall i \neq j$

$$d(\Sigma_i, \Sigma_j) = \|\Sigma_i - \Sigma_j\| \geq \lambda\sqrt{k-1} \tag{4.108}$$

We have $\alpha_r = \lambda\sqrt{k-1}$.

Now we derive an upper bound for the KL divergence

$$D(P_j\|P_i) = \frac{n}{2}[tr(\Sigma_j\Sigma_i^{-1}) - \log\det(\Sigma_j\Sigma_i^{-1}) - p] \tag{4.109}$$

Note $\det(\Sigma_i) = \det(\Sigma_j)$ since Σ_i is just a reordering of Σ_j. Thus $\log\det(\Sigma_j\Sigma_i^{-1}) = 0$. Also note $\Sigma_i^{-1} = I_{p\times p} - yB_i + (x-1)D_i$ with

$$x = \frac{1 + \lambda(k-2)}{1 + \lambda(k-2) - \lambda^2(k-1)} \tag{4.110}$$

$$y = \frac{\lambda}{1 + \lambda(k-2) - \lambda^2(k-1)} \tag{4.111}$$

Now suppose $|B_i \cap B_j| = s$, we have

$$\begin{aligned} tr(\Sigma_j \Sigma_i^{-1}) &= x(k-s) + (x - \lambda y(s-1))s + k - s + p - (2k - s) && (4.112)\\ &= p - k + xk - \lambda y s(s-1) && (4.113)\\ &\leq p + k(x-1) && (4.114) \end{aligned}$$

Hence

$$\begin{aligned} D(P_j \| P_i) &\leq \frac{n}{2} k(x-1) && (4.115)\\ &= \frac{n\lambda^2 k(k-1)}{2(1 + \lambda(k-2) - \lambda^2(k-1))} && (4.116)\\ &= \beta_r && (4.117) \end{aligned}$$

By the generalized Fano's Lemma

$$\begin{aligned} \max_i E_{P_i}[d(\hat{\theta}, \theta(P_i))] &\geq \frac{\alpha_r}{2}(1 - \frac{\beta_r + \log 2}{\log r}) && (4.118)\\ &= \frac{\lambda\sqrt{k-1}}{2}(1 - \frac{\frac{n\lambda^2 k(k-1)}{2(1+\lambda(k-2)-\lambda^2(k-1))} + \log 2}{\log \binom{p}{k}}) && (4.119)\\ &\geq \frac{\lambda\sqrt{k}}{2}(1 - c' \frac{n\lambda^2 k^2}{(1 + \lambda k - \lambda^2 k) k \log p}) && (4.120) \end{aligned}$$

Note $\frac{1}{1+\lambda k - \lambda^2 k} < 1$. $\lambda = \frac{C}{\sqrt{n}}$ implies

$$\frac{n\lambda^2 k}{(1 + \lambda k - \lambda^2 k)\log p} \leq \frac{C^2 k}{\log p} \tag{4.121}$$

$E[\|.\|^2] \geq E[\|.\|]^2$ and the block size is upper bounded by Mk implies that if $C < \sqrt{\frac{\log p}{2k}}$, then

$$\inf_{\tilde{\Sigma}} \sup_{\Sigma \in \mathcal{F}(\lambda,k,m,M,\epsilon,\delta)} E[\|\tilde{\Sigma} - \Sigma\|^2] \geq (\frac{\lambda\sqrt{Mk}}{4})^2 \geq \frac{C^2 Mk}{16n} \tag{4.122}$$

□

4.4 Algorithm Correctness and Computational Complexity

The last theorem of this chapter analyzes success probability and computational complexity of algorithm $FBAll()$.

Theorem 4.4.1. *Suppose X satisfies Assumption 1 and Assumption 2, with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$. If $\lambda = \frac{C}{\sqrt{n}}, k \geq \log p, \frac{n}{k^2} > \frac{4}{\rho_2 d_2^2}$, and if*

$$M < \frac{1 - 3\max(\epsilon, \delta)}{2\max(\epsilon, \delta)} \tag{4.123}$$

$$C > \max(\frac{8}{\sqrt{\rho_2}((\frac{1-\delta}{2})^2 - \frac{\epsilon}{2})}, 4(M+1)\sqrt{\frac{\log 2}{(1-\delta)\rho}}) \tag{4.124}$$

then there exists γ_1, γ_2, η s.t. $FBAll(A, k, s = \eta k - 1, \theta_b = 1 - \epsilon - \gamma_2, \theta_r = 1 - \gamma_1, t)$ with input matrix $A = (\mathbb{I}(|\hat{\Sigma}_{ij}| > \frac{\lambda}{2}))_{1 \leq i,j \leq p}$ has output $\{\hat{J}_l\}_{l=1}^{m'}$ satisfying

$$P(\{\hat{J}_l\}_{l=1}^{m'} \neq \{J_l\}_{l=1}^{m}) \leq mpe^{-[\rho n(\gamma_1\eta - \max(\delta,\epsilon))\lambda^2/8 - \log 2(\eta+M)]k} \tag{4.125}$$

$$+ p^2 e^{-[(\rho_2 n\gamma_2^2\lambda^2/16 - \log 2)k - 2\log p]k} \tag{4.126}$$

$$\to 0 \tag{4.127}$$

Furthermore, for probability at least

$$1 - \frac{2}{p-1} \to 1 \tag{4.128}$$

the worst case computational complexity of the algorithm $FBAll()$ is

$$O(2^s(s\log p)^{s/2}kp^{s+2}e^{-(\rho C^2/8 - \log 2)s(s+1)/4})$$

with $s = \eta k - 1$.

Remark: In general η is a constant, and $s = \eta k = O(\log p)$. This makes the algorithm running in exponential time asymptotically. However the constant C has to be large enough to make the problem statistically identifiable, and that makes the algorithm very efficient in practice for a reasonably large range of p. For p equal to several thousands, s is usually 0 or 1. For the case $s = 0$, the computation complexity is essentially $O(p^2k)$. We will illustrate this point in simulation experiments in next chapter.

Lemma 4.4.2. *Suppose X satisfies Assumption 1 with covariance matrix Σ,*

$$\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$$

For $t \in (0, \lambda)$, let

$$H_1(t, \gamma, \eta) = \{\exists J_l, \exists i \notin J_l, \exists J \subset J_l \text{ s.t. } |J| \geq \eta k, \frac{1}{|J|}\sum_{j \in J}\mathbb{I}(|\hat{\Sigma}_{ij}| > t) > \gamma\} \tag{4.129}$$

$$H_2(t, \gamma, \eta) = \{\exists J_l, \exists i \in J_l, \exists J \subset J_l \text{ s.t. } |J| \geq \eta k, \frac{1}{|J|}\sum_{j \in J}\mathbb{I}(|\hat{\Sigma}_{ij}| < t) > \gamma\} \tag{4.130}$$

if $\gamma\eta > \max(\delta, \epsilon)$, $t \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma\eta-\epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma\eta-\delta)k/2}}))$,

$$P(H_1(t,\gamma,\eta) \cup H_2(t,\gamma,\eta)) \leq \sum_{l=1}^{m} p2^{|J_l|} \max_{J \subset J_l} 2^{|J|} e^{-\rho n(\gamma|J| - \max(\delta,\epsilon)k)\min(t^2,(\lambda-t)^2)/2} \tag{4.131}$$

and if $\frac{\eta}{\eta+M}\rho n\gamma \min(t^2, (\lambda-t)^2)/2 > \log 2$,

$$P(H_1(t,\gamma,\eta) \cup H_2(t,\gamma,\eta)) < mpe^{-[\rho n(\gamma\eta - \max(\delta,\epsilon))\min(t^2,(\lambda-t)^2)/2 - \log 2(\eta+M)]k} \tag{4.132}$$

Proof. Similar to the proof of Lemma 4.2.4, for $\forall t < \frac{d}{\sqrt{(\gamma|J|-\delta k)/2}} \leq \frac{d}{\sqrt{(\gamma\eta-\delta)k/2}}$

$$P(H_1(t,\gamma,\eta)) \leq \sum_{l=1}^{m} \sum_{i \notin J_l} \sum_{J \subset J_l} P(\mathcal{C}_i(t,J) > \gamma - \frac{\delta k}{|J|}) \tag{4.133}$$

$$\leq \sum_{l=1}^{m} (p - |J_l|) 2^{|J_l|} \max_{J \subset J_l} 2^{|J|} e^{-\rho n(\gamma|J| - \delta k)t^2/2} \tag{4.134}$$

and if $\lambda - t < \frac{d}{\sqrt{(\gamma|J|-\epsilon k)/2}} \leq \frac{d}{\sqrt{(\gamma\eta-\epsilon)k/2}}$,

$$P(H_2(t,\gamma,\eta)) \leq \sum_{l=1}^{m} \sum_{i \in J_l} \sum_{J \subset J_l} P(\mathcal{C}_i(\lambda - t, J) > \gamma - \frac{\epsilon k}{|J|}) \tag{4.135}$$

$$\leq \sum_{l=1}^{m} |J_l| 2^{|J_l|} \max_{J \subset J_l} 2^{|J|} e^{-\rho n(\gamma|J| - \epsilon k)(\lambda-t)^2/2} \tag{4.136}$$

Thus if $\gamma\eta > \max(\delta, \epsilon)$, $t \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma\eta-\epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma\eta-\delta)k/2}}))$,

$$P(H_1(t,\gamma,\eta) \cup H_2(t,\gamma,\eta)) \leq \sum_{l=1}^{m} p2^{|J_l|} \max_{J \subset J_l} 2^{|J|} e^{-\rho n(\gamma|J| - \max(\delta,\epsilon)k)\min(t^2,(\lambda-t)^2)/2} \tag{4.137}$$

if $\rho n\gamma \min(t^2, (\lambda-t)^2)/2 > \log 2\frac{\eta+M}{\eta}$,

$$P(H_1(t,\gamma,\eta) \cup H_2(t,\gamma,\eta)) < mpe^{-[\rho n(\gamma\eta - \max(\delta,\epsilon))\min(t^2,(\lambda-t)^2)/2 - \log 2(\eta+M)]k} \tag{4.138}$$

□

Lemma 4.4.3. *Suppose X satisfies Assumption 1 and Assumption 2 with covariance matrix $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$. Let*

$$G(t,\gamma,s,w) = \{\exists J, L \ s.t. \ |J| \geq s, |L| \geq w, \frac{1}{|J||L|} \sum_{j \in J, l \in L} \mathbb{I}(|\hat{\Sigma}_{jl} - \Sigma_{jl}| > t) > \gamma\} \tag{4.139}$$

if $(\rho_2 n\gamma^2 t^2/4 - \log 2)\min(s,w) > 2\log p$, *for* $\forall t < \frac{2d_2}{\gamma\sqrt{|J||L|}}$,

$$P(G(t,\gamma,s,w)) \leq p^2 e^{-[(\rho_2 n\gamma^2 t^2/4 - \log 2)\min(s,w) - 2\log p]\max(s,w)} \tag{4.140}$$

Proof. Similar to the proof of Lemma 4.2.5, if

$$(\rho_2 n\gamma^2 t^2/4 - \log 2)\min(s,w) > 2\log p$$

for $\forall t < \frac{2d_2}{\gamma\sqrt{|J||L|}}$,

$$P(G(t,\gamma,s,w)) \leq \sum_{|J|=s}^{p}\sum_{|L|=w}^{p}\binom{p}{|J|}\binom{p}{|L|}P(\frac{1}{\sqrt{|J||L|}}\sum_{j\in J,l\in L}\mathbb{I}(|\hat{\Sigma}_{jl}-\Sigma_{jl}|>t) > \gamma\sqrt{|J||L|}) \tag{4.141}$$

$$\leq \sum_{|J|=s}^{p}\sum_{|L|=w}^{p} p^{|J|+|L|}2^{|J||L|}e^{-\rho_2 n|J||L|\gamma^2t^2/4} \tag{4.142}$$

$$\leq p^2 e^{-[(\rho_2 n\gamma^2t^2/4-\log 2)\min(s,w)-2\log p]\max(s,w)} \tag{4.143}$$

□

Lemma 4.4.4. *Define: A zero-mean random variable Y is sub-exponential if $\exists d > 0$ s.t.*

$$E[e^{tY}] \leq \infty \text{ for } \forall|t| \leq d \tag{4.144}$$

Claim: Y is sub-exponential if and only if $\exists\rho', d'$ s.t. $E[\exp(tY)] \leq \exp(\frac{t^2\rho'}{2})$ for $\forall|t| < d'$.

Proof. For t close to 0,

$$E[e^{tY}] = 1 + \frac{t^2E[Y^2]}{2} + o(t^2) \tag{4.145}$$

$$e^{\frac{t^2\rho'}{2}} = 1 + \frac{t^2\rho'}{2} + o(t^2) \tag{4.146}$$

Hence if $\rho' > E[Y^2]$, then $\exists d'$ s.t. $E[\exp(tY)] \leq \exp(\frac{t^2\rho'}{2})$ for $\forall|t| < d'$. □

Lemma 4.4.5. *Suppose X satisfies Assumption 1 and Assumption 2. Let*

$$\mathcal{D}(t,J,L) = \frac{1}{|J|}\sum_{j\in J}\prod_{i\in L}\mathbb{I}(|\hat{\Sigma}_{ij}-\Sigma_{ij}|>t) \tag{4.147}$$

Then $\exists d_3 > 0$ s.t. for $\forall|t'| < d_3$,

$$P(|\mathcal{D}(t,J,L) - E[\mathcal{D}(t,J,L)]| > t') < 2e^{-\frac{t'^2}{2}e^{(\rho n t^2/2-\log 2)|L|}} \tag{4.148}$$

Furthermore, let

$$\mathcal{G}(t',t,s) = \{\exists L \text{ s.t. } |L| = s, |\mathcal{D}(t,[p],L) - E[\mathcal{D}(t,[p],L)]| > t'\} \tag{4.149}$$

then

$$P(\mathcal{G}(t',t,s)) < 2p^s e^{-\frac{t'^2}{2}e^{(\rho n t^2/2-\log 2)s}} \tag{4.150}$$

Proof. For $\forall t < \frac{\sqrt{2}d}{\sqrt{|L|}}$, Assumption 1 implies

$$E[\mathcal{D}(t,J,L)] \leq \max_{j\in J} E[\prod_{i\in L} \mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t)] \tag{4.151}$$

$$= \max_{j\in J} P(\sum_{i\in L} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > |L|t) \tag{4.152}$$

$$< e^{-(\rho n t^2/2 - \log 2)|L|} \tag{4.153}$$

Similarly

$$E[\mathcal{D}(t,J,L)^2] < \max_{j_1,j_2\in J} E[\prod_{i\in L} \mathbb{I}(|\hat{\Sigma}_{ij_1} - \Sigma_{ij_1}| > t) \prod_{i\in L} \mathbb{I}(|\hat{\Sigma}_{ij_2} - \Sigma_{ij_2}| > t)] \tag{4.154}$$

$$= \max_{j\in J} P(\sum_{i\in L} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > |L|t) \tag{4.155}$$

$$< e^{-(\rho n t^2/2 - \log 2)|L|} \tag{4.156}$$

Let $Y = \mathcal{D}(t,J,L) - E[\mathcal{D}(t,J,L)]$. Generalization of Hölder's inequality implies

$$E[\exp(\tau Y)] = E[\exp(\tau \mathcal{D}(t,J,L) - \tau E[\mathcal{D}(t,J,L)])] \tag{4.157}$$

$$= \exp(-\tau E[\mathcal{D}(t,J,L)]) E[\exp(\tau \mathcal{D}(t,J,L))] \tag{4.158}$$

$$\leq E[\prod_{j=J} \exp(\frac{\tau}{|J|} \prod_{i\in L} \mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t))] \tag{4.159}$$

$$\leq \prod_{j\in J} E[\exp(\tau \prod_{i\in L} \mathbb{I}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t))]^{\frac{1}{|J|}} \tag{4.160}$$

$$\leq \max_{j\in J} e^{\tau} P(\sum_{i\in L} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > |L|t) + 1 \tag{4.161}$$

$$< \infty \tag{4.162}$$

Hence Y is sub-exponential and $E[Y^2] < E[\mathcal{D}(t,J,L)^2] < e^{-(\rho n t^2/2 - \log 2)|L|}$, Lemma 4.4.4 implies that $\exists d_3 > 0$ s.t. for $\forall |t'| < d_3$,

$$P(|\mathcal{D}(t,J,L) - E[\mathcal{D}(t,J,L)]| > t') = P(|Y| > t') < 2e^{-\frac{t'^2}{2} e^{(\rho n t^2/2 - \log 2)|L|}} \tag{4.163}$$

Union bound implies the upper bound of $P(\mathcal{G}(t',t,s))$. □

Proof of Theorem 4.4.1:

Proof. Suppose $H_1(\frac{\lambda}{2},\gamma_1,\eta)^c \cap H_2(\frac{\lambda}{2},\gamma_1,\eta)^c \cap G(\frac{\lambda}{2},\gamma_2,k,k)^c$ happens with γ_1,γ_2 to be chosen later. Recall the input matrix $A = (\mathbb{I}(|\hat{\Sigma}_{ij}| > \frac{\lambda}{2}))_{1\leq i,j\leq p}$. Denote $\{\hat{J}_l\}_{l=1}^{m'}$ the output of

algorithm $FBAll(A, k, s, \theta_b, \theta_r, t)$. Each $\hat{J}_l$ is an output of $FBRec(A, L, \emptyset, k, s, \theta_b, \theta_r, t)$ with $L = [p]$ initially. By construction of the algorithm, $\hat{J}_l$ satisfies

$$|\hat{J}_l| \geq k \tag{4.164}$$

$$\frac{1}{|\hat{J}_l|} \sum_{j \in \hat{J}_l} A_{ij} \geq \theta_r \text{ for } \forall i \in \hat{J}_l \tag{4.165}$$

$$\frac{1}{|\hat{J}_l|^2} \sum_{i,j \in \hat{J}_l} A_{ij} \geq \theta_b \tag{4.166}$$

Similar to the proof of Lemma 4.2.6, if following holds:

$$\frac{1-\delta}{2} - M\gamma_1 \geq (1 - \theta_r) \tag{4.167}$$

$$2(\frac{1-\delta}{2})^2 - \gamma_2 \geq 1 - \theta_b \tag{4.168}$$

$$1 - \gamma_1 \geq \theta_r \tag{4.169}$$

$$\max(1 - \gamma_1, 1 - \epsilon - \gamma_2) \geq \theta_b \tag{4.170}$$

then $H_1(\frac{\lambda}{2}, \gamma_1, \eta)^c \cap H_2(\frac{\lambda}{2}, \gamma_1, \eta)^c \cap G(\frac{\lambda}{2}, \gamma_2, k, k)^c$ implies that there exists some l' s.t. $\hat{J}_l \subset J_{l'} = \mathbb{J}(B_{l'})$. Thus if

$$\theta_r = 1 - \gamma_1 \tag{4.171}$$

$$\theta_b = \max(1 - \gamma_1, 1 - \epsilon - \gamma_2) \tag{4.172}$$

$$\gamma_1 \leq \frac{1-\delta}{2(M+1)} \tag{4.173}$$

$$\gamma_2 \leq \max((\frac{1-\delta}{2})^2 - \frac{\epsilon}{2}, 2(\frac{1-\delta}{2})^2 - \gamma_1) \tag{4.174}$$

then $FBRec()$ does not recover any wrong block and always recovers correct sub-blocks. It remains to show that it indeed finds all the blocks with full size.

Calling $FBAll(A, k, s, \theta_b, \theta_r, t)$ would eventually call $FB(A, J, k, t)$ with $|J| = s + 1$. Suppose $J \subset J_l = \mathbb{J}(B_l)$ for some l, and $J \cap \cup_{i \neq l} J_i = \emptyset$, which at some step always happens if $s < (1 - \delta)k - 1$ and since $FBAll()$ exhaustively search over all J satisfying

$$A_{ij} \neq 0 \text{ for } \forall i \neq j \in J \tag{4.175}$$

If $s + 1 \geq \eta k$, then $H_1(\frac{\lambda}{2}, \gamma_1, \eta)^c \cap H_2(\frac{\lambda}{2}, \gamma_1, \eta)^c$ implies that

$$\sum_{j \in J} \mathbb{I}(|\hat{\Sigma}_{ij}| > \frac{\lambda}{2}) > (1 - \gamma_1)|J| \text{ for } \forall i \in J \tag{4.176}$$

$$\sum_{j \in J} \mathbb{I}(|\hat{\Sigma}_{ij}| > \frac{\lambda}{2}) < \gamma_1|J| \text{ for } \forall i \notin J \tag{4.177}$$

If $\gamma_1 < \frac{1}{2}$, then the output J_0 of $FB(A, J, k, t)$ satisfies $|J_0| = k$ and $J_0 \subset J_l$. By the same reasoning, $FBS(A, J_0, \theta_b, \theta_r, t)$ recovers the full size of J_l with θ_r, θ_b defined in 4.171 and 4.172. This shows the correctness of the algorithm if event $H_1(\frac{\lambda}{2}, \gamma_1, \eta)^c \cap H_2(\frac{\lambda}{2}, \gamma_1, \eta)^c \cap G(\frac{\lambda}{2}, \gamma_2, k, k)^c$ happens.

To bound the error probability of $FBAll()$, Lemma 4.4.2 and Lemma 4.4.3 implies that if

$$\gamma_1\eta > \max(\delta, \epsilon) \tag{4.178}$$

$$\frac{\lambda}{2} \in (\max(0, \lambda - \frac{d}{\sqrt{(\gamma_1\eta - \epsilon)k/2}}), \min(\lambda, \frac{d}{\sqrt{(\gamma_1\eta - \delta)k/2}})) \tag{4.179}$$

$$\log 2 < \frac{\eta}{\eta + M}\rho n\gamma_1\lambda^2/8 \tag{4.180}$$

$$\frac{\log p}{k} < \rho_2 n\gamma_2^2\lambda^2/32 - \log 2/2 \tag{4.181}$$

$$\frac{\lambda}{2} < \frac{2d_2}{\gamma_2 k} \tag{4.182}$$

then

$$P(H_1(\frac{\lambda}{2}, \gamma_1, \eta) \cup H_2(\frac{\lambda}{2}, \gamma_1, \eta) \cup G(\frac{\lambda}{2}, \gamma_2, k, k)) \leq mpe^{-[\rho n(\gamma_1\eta - \max(\delta,\epsilon))\lambda^2/8 - \log 2(\eta+M)]k} \tag{4.183}$$

$$+ p^2 e^{-[(\rho_2 n\gamma_2^2\lambda^2/16 - \log 2)k - 2\log p]k} \tag{4.184}$$

$$\to 0 \tag{4.185}$$

Combining 4.171 – 4.174 and 4.178 – 4.182, and $\lambda = \frac{C}{\sqrt{n}}$, $k \geq \log p$, we have

$$\theta_r = 1 - \gamma_1 \tag{4.186}$$

$$\theta_b = \max(1 - \gamma_1, 1 - \epsilon - \gamma_2) \tag{4.187}$$

$$\max(\frac{\max(\epsilon, \delta)}{\eta}, \frac{8\log 2(\eta + M)}{\rho C^2\eta}) < \gamma_1 \leq \frac{1 - \delta}{2(M + 1)} \tag{4.188}$$

$$\frac{8}{\sqrt{\rho_2}C} < \gamma_2 \leq \min(\frac{4d_2\sqrt{n}}{Ck}, (\frac{1 - \delta}{2})^2 - \frac{\epsilon}{2}) \tag{4.189}$$

where 4.188 and 4.189 are equivalent to

$$\eta > \frac{2(M + 1)\max(\epsilon, \delta)}{1 - \delta} \tag{4.190}$$

$$\eta > \frac{M(M + 1)16\log 2}{\rho C^2(1 - \delta) - (M + 1)16\log 2} \tag{4.191}$$

$$\frac{\sqrt{n}}{k} > \frac{2}{d_2\sqrt{\rho_2}} \tag{4.192}$$

$$C > \frac{8}{\sqrt{\rho_2}((\frac{1-\delta}{2})^2 - \frac{\epsilon}{2})} \tag{4.193}$$

Thus if $\lambda = \frac{C}{\sqrt{n}}, k \geq \log p, \frac{n}{k^2} > \frac{4}{\rho_2 d_2^2}$, and if

$$M < \frac{1 - 3\max(\epsilon, \delta)}{2\max(\epsilon, \delta)} \tag{4.194}$$

$$C > \max(\frac{8}{\sqrt{\rho_2}((\frac{1-\delta}{2})^2 - \frac{\epsilon}{2})}, 4(M+1)\sqrt{\frac{\log 2}{(1-\delta)\rho}}) \tag{4.195}$$

then there exists γ_1, γ_2, η s.t. $FBAll(A, k, s = \eta k - 1, \theta_b = 1 - \epsilon - \gamma_2, \theta_r = 1 - \gamma_1, t)$ perfectly recovers $\{J_l\}_{l=1}^m$ with probability going to 1.

Next we calculate the computation complexity of $FBRec()$. Suppose $\cap_{i=1}^s \mathcal{G}(t_i', \frac{\lambda}{2}, i)^c$ happens with t_i' satisfying

$$t_i' = 2\sqrt{i \log p} e^{-(\rho n \lambda^2/8 - \log 2)i/2} \tag{4.196}$$

$$= 2\sqrt{i \log p} \Delta^i \tag{4.197}$$

where $\Delta = e^{-(\rho n \lambda^2/8 - \log 2)/2}$. Recall that $FBAll()$ searches over all $J \in \mathcal{J}(s+1)$ defined as

$$\mathcal{J}(s+1) = \{J : |J| = s+1,\ A_{ij} \neq 0 \text{ for } \forall i \neq j \in J\} \tag{4.198}$$

For each J, $FB()$ takes $O(pk)$. Hence $FBAll()$ takes $O(|\mathcal{J}(s+1)|pk)$ with

$$|\mathcal{J}(s+1)| = p \prod_{i=1}^{s} (p t_i') \tag{4.199}$$

$$\leq 2^s (s \log p)^{s/2} p^{s+1} \Delta^{s(s+1)/2} \tag{4.200}$$

Hence the overall worst case computational complexity for $FBAll()$ is

$$O(2^s (s \log p)^{s/2} k p^{s+2} \Delta^{s(s+1)/2})$$

For the case $s = 0$, the dominant term is $O(kp^2)$. Furthermore, this is true with probability at least

$$P(\cap_{i=1}^s \mathcal{G}(t_i', \frac{\lambda}{2}, i)^c) \geq 1 - 2\sum_{i=1}^{s} p^i e^{-\frac{t_i'^2}{2}} e^{(\rho n t^2/2 - \log 2)i} \tag{4.201}$$

$$\geq 1 - 2\sum_{i=1}^{s} p^{-i} \tag{4.202}$$

$$\geq 1 - \frac{2}{p-1} \to 1 \tag{4.203}$$

□

Chapter 5

Experiment

5.1 Single Block Spike model

Block Recovery

In this section we simulate experiment for single block recovery. Suppose $X \in \mathbb{R}^{n\times p}$ with i.i.d. $X_{(i)} \sim N(0,\Sigma)$ are observed. After unknown permutation of indices, Σ belongs to $\mathcal{E}_{\beta,k}$:

$$\mathcal{E}_{\beta,k} = \{\Sigma : \Sigma = \beta zz^t + \begin{bmatrix} \frac{k-1}{k} I_k & 0 \\ 0 & I_{p-k} \end{bmatrix}, z_i = \pm\frac{1}{\sqrt{k}}\} \tag{5.1}$$

Note that $\mathcal{E}_{\beta,k} \subset \mathcal{F}(\frac{\beta}{k}, 1, k, 1, 0, 0)$, i.e. it is a single block special case of our multiple block model. Also notice that it is slightly different from the spike model $\mathcal{E}_\beta$ in 2.11 used in Amini and Wainwright (AW2009-1). Besides this difference, we follow their experiment setting: for each given p, fix $\beta = 3$ to be constant, let $k = 3\log p$, and let n scales with the signal to noise ratio $\frac{n}{k\log p}$, which increases from 1 to 5.

Suppose k and $\lambda = \frac{\beta}{k}$ are known, the goal is to recover $supp(z)$. The metric of evaluation is recovery proportion(RP):

$$RP = \frac{\text{number of correct variables recovered}}{k} \tag{5.2}$$

$$= 1 - \frac{\text{Hamming distance between estimator and } supp(z)}{k} \tag{5.3}$$

$RP = 1$ is perfect recovery and $RP = 0$ is the worst possible. Given threshold $h = \frac{\beta}{2k}$, two input matrices $A = E_h$ and $A = W_h$ for $FBRec()$ are used,

$$E_h = (\mathbb{I}(|\hat{\Sigma}_{ij}| > h))_{1\le i,j\le p} \tag{5.4}$$

$$W_h = (|\hat{\Sigma}_{ij}|\mathbb{I}(|\hat{\Sigma}_{ij} > h| \text{ and } i \neq j))_{1\le i,j\le p} \tag{5.5}$$

Performances of following 3 algorithms are compared based on simulation:

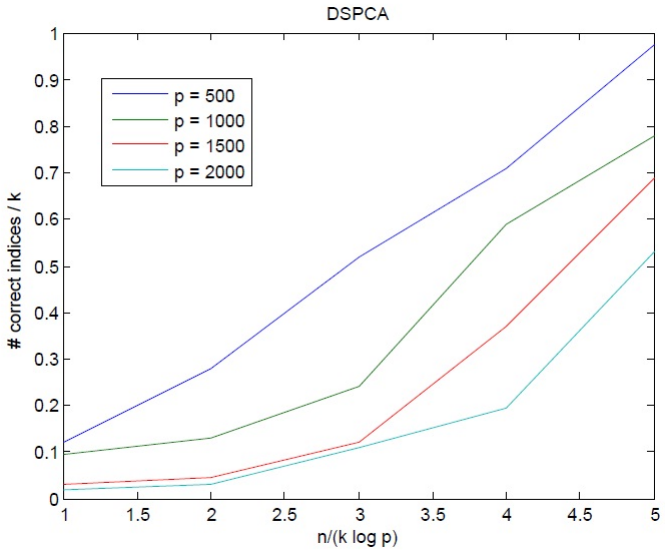

Figure 5.1: DSPCA: 20 iterations average RP with input matrix $\hat{\Sigma}$

- DSPCA: Use $\hat{\Sigma}$ as input matrix. Take eigen-decomposition of the output matrix of DSPCA, pick top k largest entries in absolute value of the eigenvector corresponding to the largest eigenvalue.
- $FBRec(0)$: $s = 0$. Use $A = E_h$ or W_h as input matrix. Let $L = [p], J = \emptyset$, k is given, θ_r, θ_b as chosen in Theorem 4.4.1, $t = 10$.
- $FBRec(1)$: $s = 1$. Use $A = E_h$ or W_h as input matrix. Let $L = [p], J = \emptyset$, k is given, θ_r, θ_b as chosen in Theorem 4.4.1, $t = 10$.

Note that $FBRec(0)$ with $s = 0$ is similar to $TPower$ by X.T. Yuan and T. Zhang (YZ2011-1). For dimension $p = 500, 1000, 1500, 2000$, average RP over 20 iterations are reported in Figure 5.1, Figure 5.2, Figure 5.3. As shown by the result, input W_h indeed is better than input E_h for both $FBRec(0)$ and $FBRec(1)$, while $FBRec(0)$ is more sensitive to dimension increment. The average running time for various algorithm is reported in Table 5.1. As the running time for $DSPCA$ does not depend on the signal to noise ratio(SNR) $\frac{n}{k \log p}$, for each dimension p the average running time over all SNR is reported. Notice that the running time for both $FBRec(0)$ and $FBRec(1)$ heavily depends on SNR, i.e. for fixed dimension p, $FBRec()$ runs faster for larger SNR.

Covariance Estimation

As in previous section, suppose the same data X is observed with known k and β. Next goal is covariance estimation in spectral norm. Denote $\tilde{\Sigma}$ as the generic estimator, the

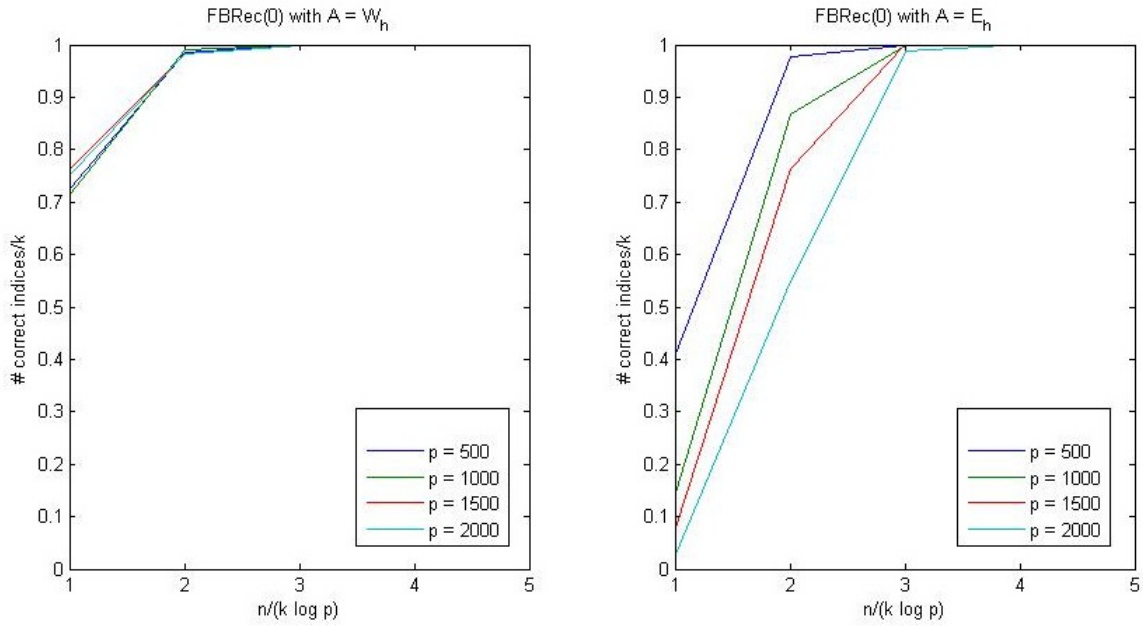

Figure 5.2: FBRec(0): 20 iterations average RP with $s = 0$, input matrix W_h and E_h

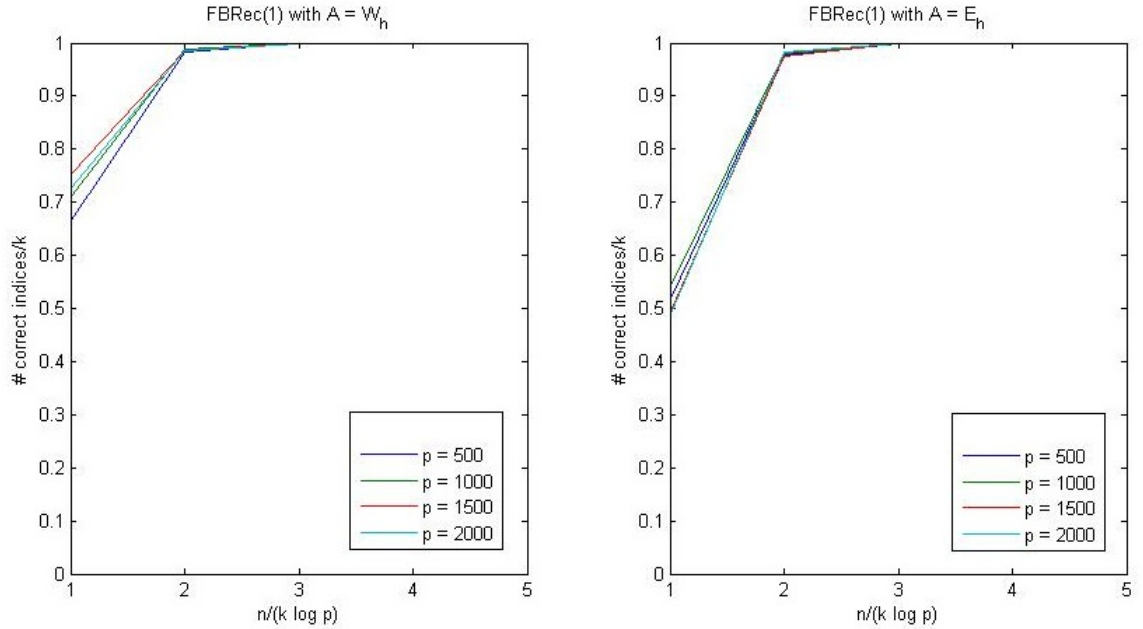

Figure 5.3: FBRec(1): 20 iterations average RP with $s = 1$, input matrix W_h and E_h

p	$\frac{n}{k \log p}$	DSPCA	FBRec(0)	FBRec(1)
500	1	16.6	0.0217	0.0805
500	2	16.6	0.0055	0.0112
500	3	16.6	0.0037	0.0045
500	4	16.6	0.0022	0.0023
500	5	16.6	0.0020	0.0020
1000	1	106.9	0.3048	0.6123
1000	2	106.9	0.0328	0.0558
1000	3	106.9	0.0153	0.0187
1000	4	106.9	0.0079	0.0080
1000	5	106.9	0.0059	0.0058
1500	1	331.1	1.2577	1.7902
1500	2	331.1	0.1038	0.1365
1500	3	331.1	0.0364	0.0383
1500	4	331.1	0.0167	0.0176
1500	5	331.1	0.0122	0.0122
2000	1	899.2	3.9335	4.4427
2000	2	899.2	0.2647	0.4375
2000	3	899.2	0.0788	0.0875
2000	4	899.2	0.0354	0.0465
2000	5	899.2	0.0199	0.0191

Table 5.1: Average run time in seconds over 20 iterations

performances is evaluated based on the spectral norm of the difference between the truth Σ and the estimator $\tilde{\Sigma}$:

$$\|\Delta\| = \|\tilde{\Sigma} - \Sigma\|$$

Following 3 estimators are compared:

- Oracle: As if the permutation is known, keep entries of $\hat{\Sigma}$ within the block and on the diagonal, and make all the other entries 0.
- Block estimate: Use $FBRec()$ to recover the block, keep entries of $\hat{\Sigma}$ within the estimated block and on the diagonal, make every other entries 0.
- Threshold: Pick threshold h that minimizes $\|\Delta\|$, then estimate Σ by $T_h(\hat{\Sigma})$.

For dimension $p = 500, 1000, 1500, 2000$, average $\|\Delta\|$ over 20 iterations for above estimators are reported in Figure 5.4.

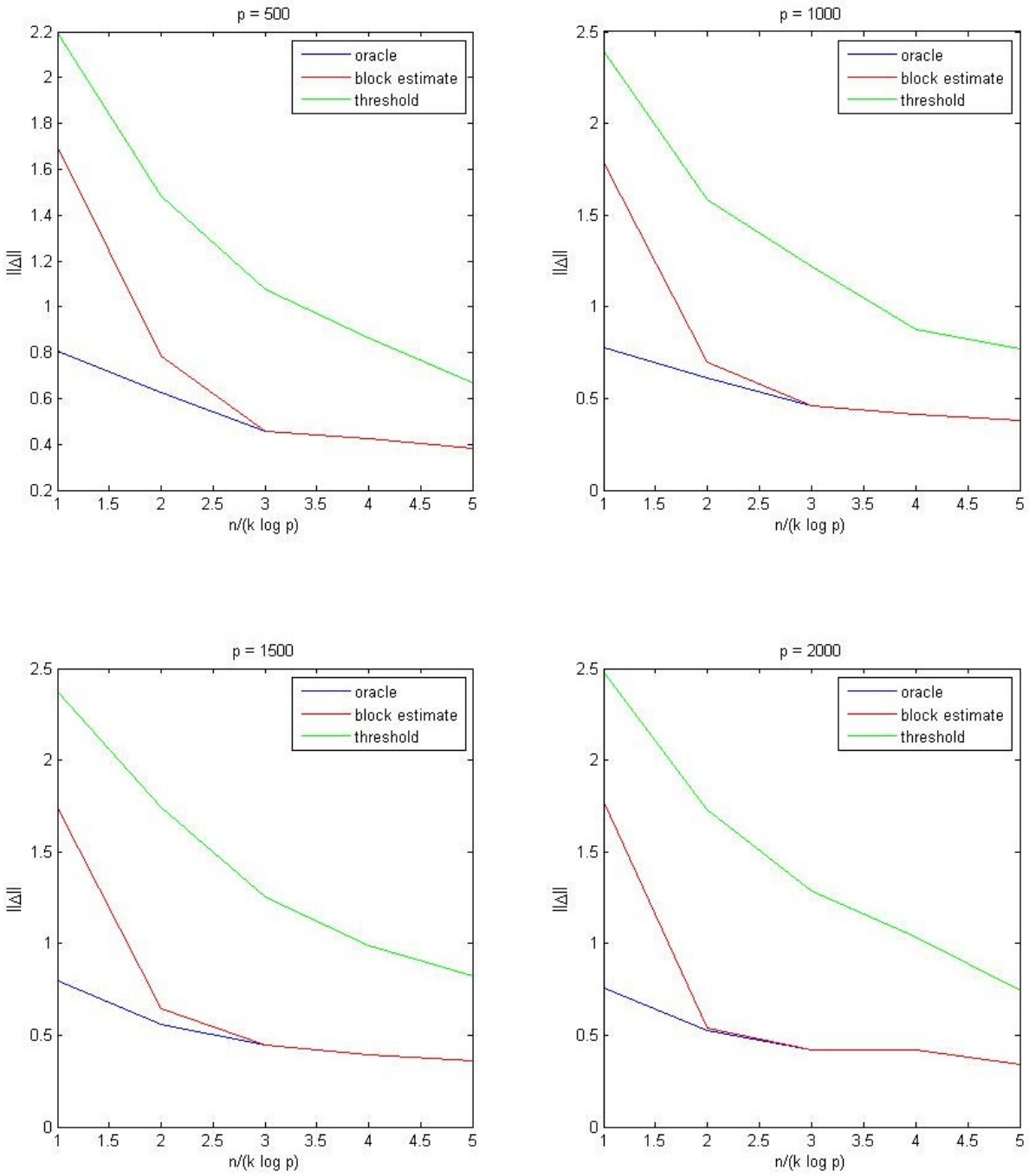

Figure 5.4: 20 iteration average $\|\Delta\|$: spectral norm of difference between the population and the estimator

$p \diagdown \frac{n}{k \log p}$	2	3	4	5
500	9.71	4.92	5.19	5.24
1000	41.69	27.89	29.41	30.44
1500	147.21	112.25	75.02	78.40
2000	248.43	156.84	158.38	165.79

Table 5.2: $FBAll()$: 30 iteration average run time in seconds. SNR scale $\frac{n}{k \log p}$.

5.2 Multiple Block Model

In this section we simulate experiments for multiple blocks recovery and covariance estimation. Suppose $X \in \mathbb{R}^{n \times p}$ with i.i.d. $X_{(i)} \sim N(0, \Sigma)$ are observed. $\Sigma \in \mathcal{F}(\lambda, k, m, M, \epsilon, \delta)$. Same as previously, consider $p = 500, 1000, 1500, 2000$. For each given p, let $k = 3 \log p, \lambda = \frac{3}{k}, M = 2, m = p/(Mk), \epsilon = \delta = 0.1$, and n scales with $\frac{n}{k \log p}$ which increases from 2 to 5. The size of each block is uniformly random from $[k, Mk] = [k, 2k]$. Given estimator $\{\hat{J}_l\}_{l=1}^{m'}$ of the true blocks $\{J_l\}_{l=1}^{m}$, the performance of support recovery is evaluated by multiple recovery proportion(MRP):

$$MRP = \frac{1}{m + m'} \left(\sum_{l=1}^{m} \max_{i=1}^{m'} \frac{|J_l \cap \hat{J}_i|}{|J_l|} + \sum_{i=1}^{m'} \max_{l=1}^{m} \frac{|J_l \cap \hat{J}_i|}{|\hat{J}_i|} \right) \tag{5.6}$$

MRP can be considered as average RP over all $\{J_l\}_{l=1}^{m}$ and $\{\hat{J}_l\}_{l=1}^{m'}$. $MRP = 1$ if and only if perfect recovery $\{J_l\}_{l=1}^{m} = \{\hat{J}_l\}_{l=1}^{m'}$ with $m = m'$. $MRP = 0$ is the worst possible. $FBAll(A, k, s, \theta_b, \theta_r, t)$ is used for support recovery, where $A = W_h$ with $h = \frac{\lambda}{2} = \frac{3}{2k}$, k is given, $s = 1$, θ_b, θ_r as in Theorem 4.4.1, $t = 10$. 30 iteration average MRP is reported in Figure 5.5. The average running time for $FBAll()$ is reported in Table 5.2. Same as previous section, the three estimators: oracle, block estimate and threshold are compared in terms of spectral norm of the difference to the truth. 30 iteration average $\|\Delta\|$ for the three estimators are reported in Figure 5.7.

Next we use different SNR scale, i.e. for each given p, let $n = 3k \log p = 9(\log p)^2, \lambda = \frac{C}{\sqrt{n}}$ with C increasing from 2 to 5, and keep everything else unchange. Again use $FBAll()$ for support recovery, with W_h and $h = \frac{\lambda}{2} = \frac{C}{2\sqrt{n}}$, and keep all other setting the same. Use the same three estimators to estimate Σ. 30 iteration average MRP for support recovery is reported in Figure 5.6. Average running time for $FBAll()$ is reported in Table 5.3. And 30 iteration average $\|\Delta\|$ is reported in Figure 5.8. As shown by both experiments, perfect block recovery implies that the block estimate agrees with the oracle for large enough signal strength.

p \ C	2	3	4	5
500	12.53	5.31	5.41	6.37
1000	73.83	27.81	29.85	34.78
1500	182.81	74.65	77.66	81.14
2000	284.54	153.46	156.37	160.53

Table 5.3: $FBAll()$: 30 iteration average run time in seconds. SNR scale $C = \lambda\sqrt{n}$.

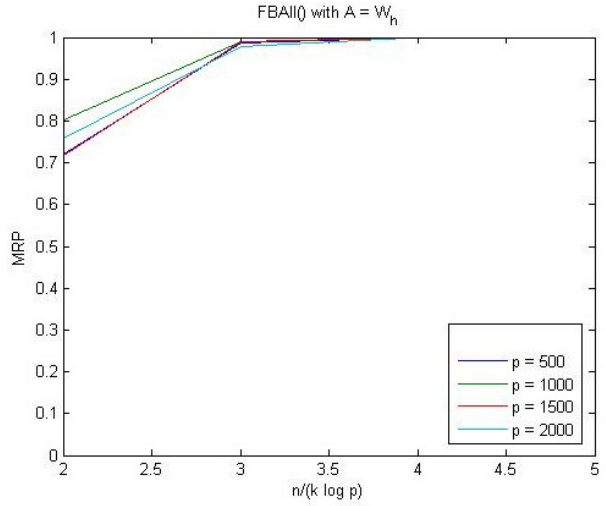

Figure 5.5: 30 iteration average MRP for $FBAll$ with input matrix W_h. SNR scale $\frac{n}{k \log p}$.

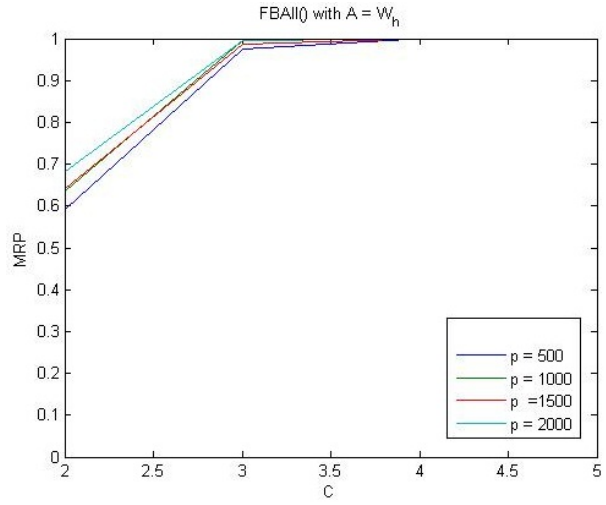

Figure 5.6: 30 iteration average MRP for $FBAll$ with input matrix W_h. SNR scale $C = \lambda\sqrt{n}$.

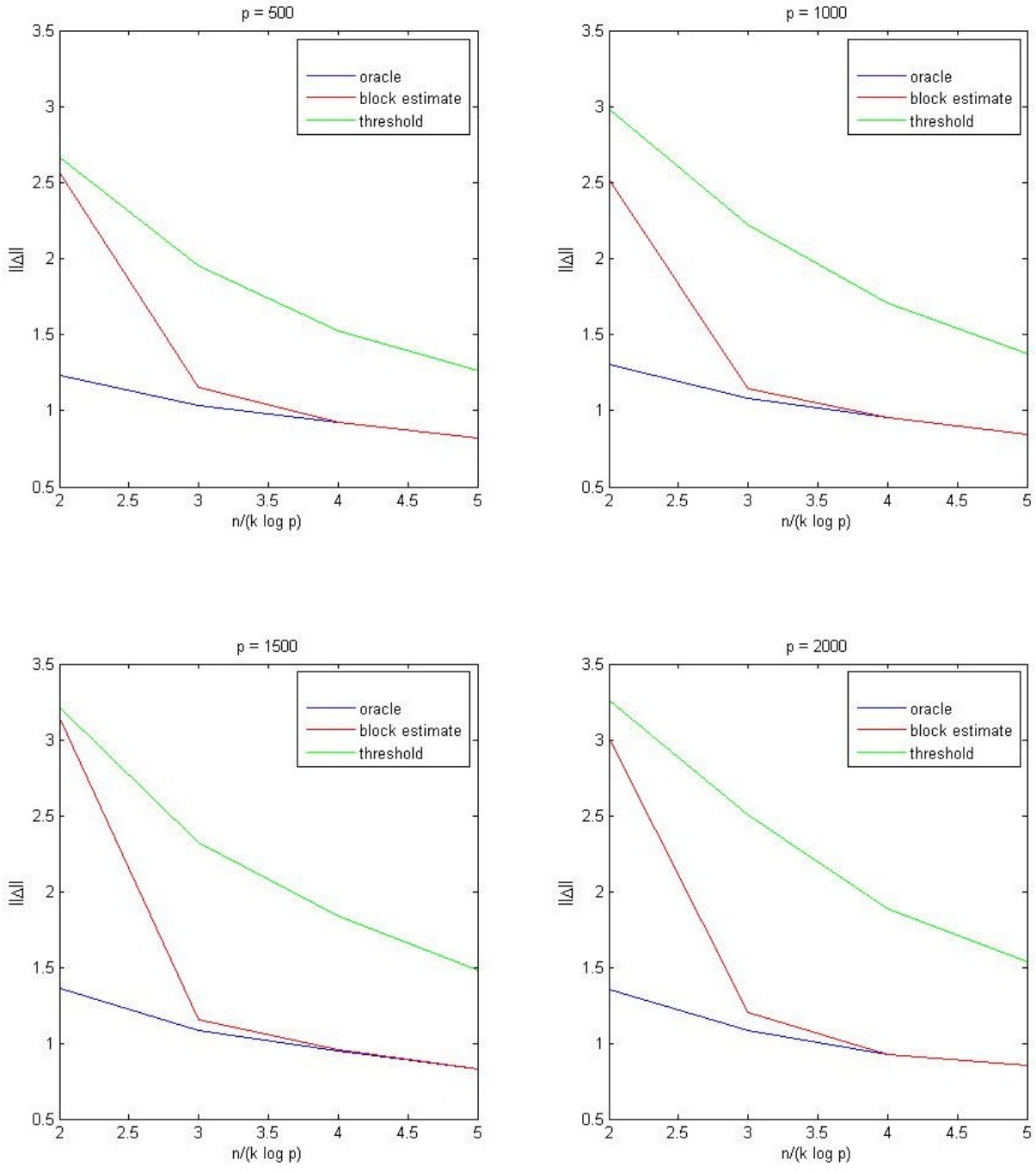

Figure 5.7: 30 iteration average $\|\Delta\|$: spectral norm of difference between the population and the estimator. SNR scale $\frac{n}{k \log p}$.

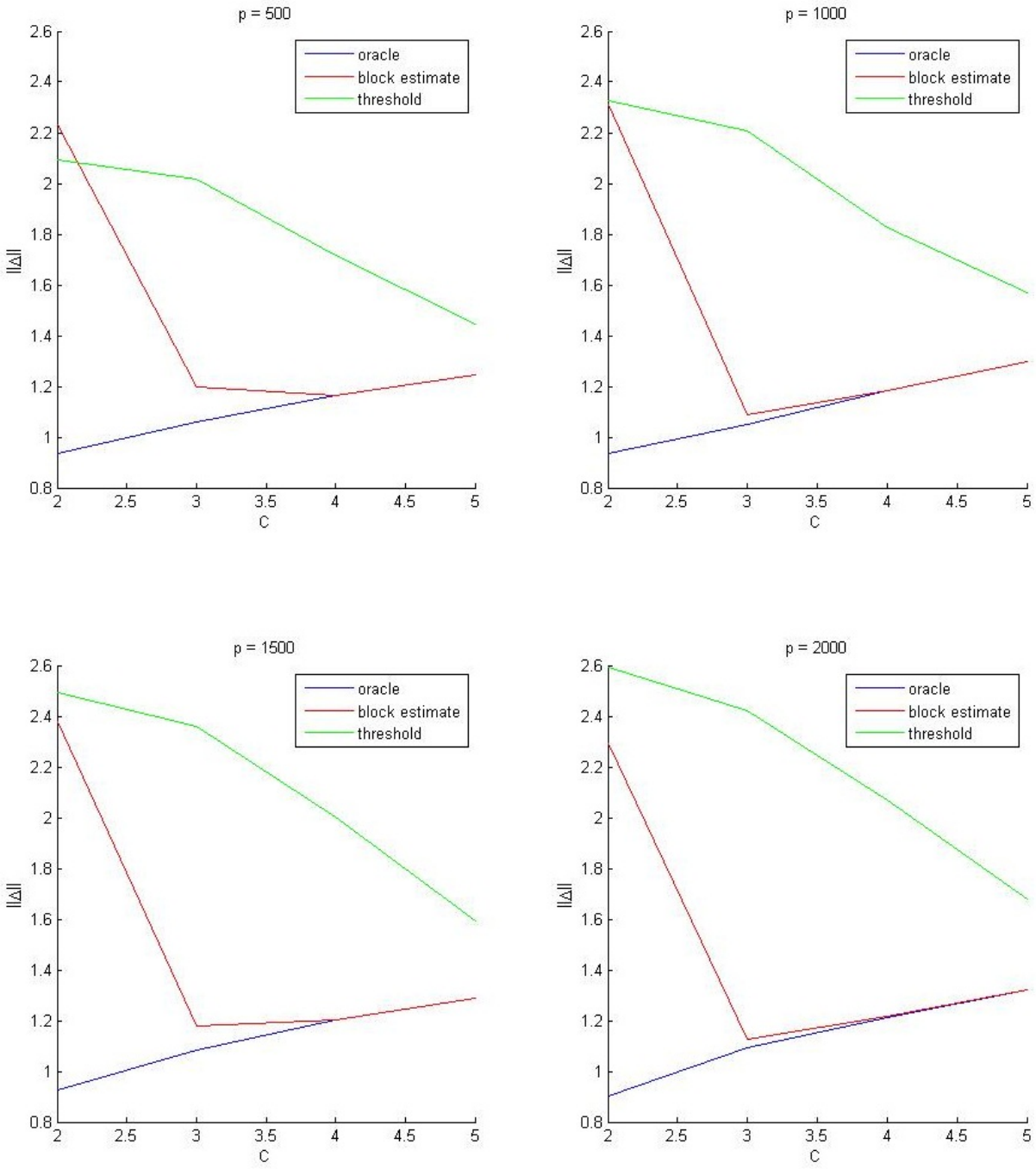

Figure 5.8: 30 iteration average $\|\Delta\|$: spectral norm of difference between the population and the estimator. SNR scale $C = \lambda\sqrt{n}$.

5.3 Application

In this section we apply the block recovery algorithm to the so called entity resolution problem to identify whether data objects from different sources represent the same entity. This problem is also known as record matching (P2002-1) (F2009-1), record linkage (N1959-1) (F1969-1) (B2004-1), or deduplication (B2003-1) (BG2004-1). Given data $B \in \mathbb{R}^{n \times p}$, each object is represented by a sparse n vector and there are p objects. $B_{ij} \geq 0$ for all i, j. The cosine similarity $A \in \mathbb{R}^{p \times p}$ is defined as

$$A_{ij} = \frac{B_i^t B_j}{\sqrt{\|B_i\|_2 \|B_j\|_2}} \tag{5.7}$$

This is following the setup in L.C. Shu et al (S2011-1).The goal is to partition indices of A into blocks such that similarity A_{ij} is large for i, j within the same block, and A_{ij} is small for i, j from different blocks. This type of data is not what we have studied up to now, but our algorithm is directly applicable. We compare following two algorithms:

- SPAN proposed by L.C. Shu et al (S2011-1): The idea is to use spectral clustering to recursively bi-partition the indices and stop while Newman-Girvan modularity is negative.
- FB: Run FBAll$(A, k, s, \theta_b, \theta_r, t)$ with parameters $k = 10, s = 0, \theta_b = 0.5, \theta_r = 0.4$.

We apply these two algorithms to a data set from Alcatel Lucent with $A \in \mathbb{R}^{p \times p}$ for $p = 3000$. This is a randomly picked subset of the data used in L.C. Shu et al (S2011-1). The original similarity matrix is plotted in Figure 5.9. The similarity matrix permuted by SPAN is plotted in Figure 5.10. The similarity matrix permuted by FB is plotted in Figure 5.11. Figure 5.12, Figure 5.13 and Figure 5.14 are a closer look to sub-matrices of similarity matrix permuted by FB.

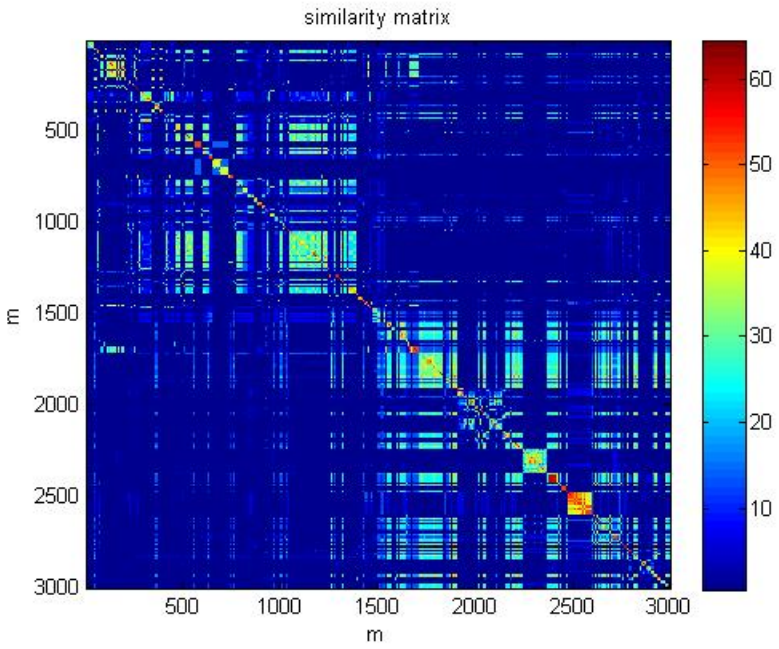

Figure 5.9: Original similarity matrix, each entry multiplied by 60

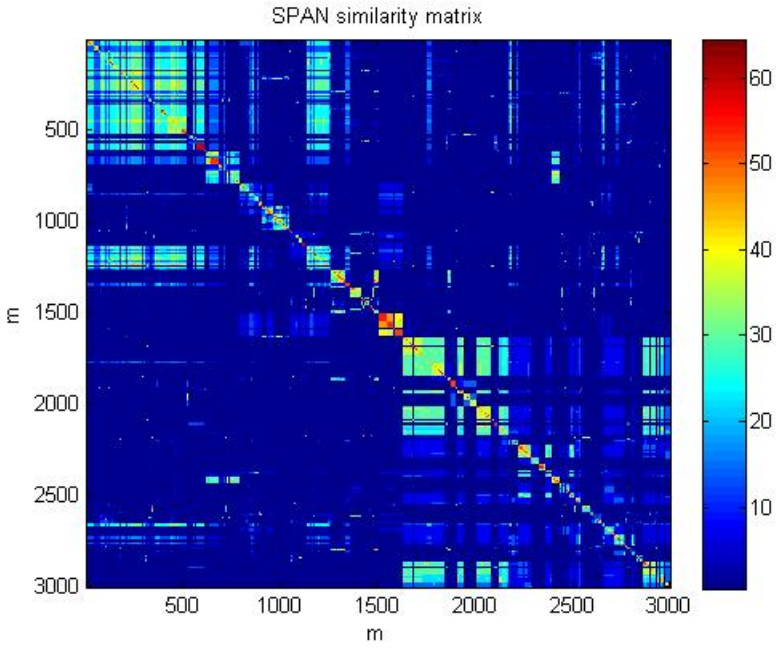

Figure 5.10: Similarity matrix permuted by SPAN, each entry multiplied by 60

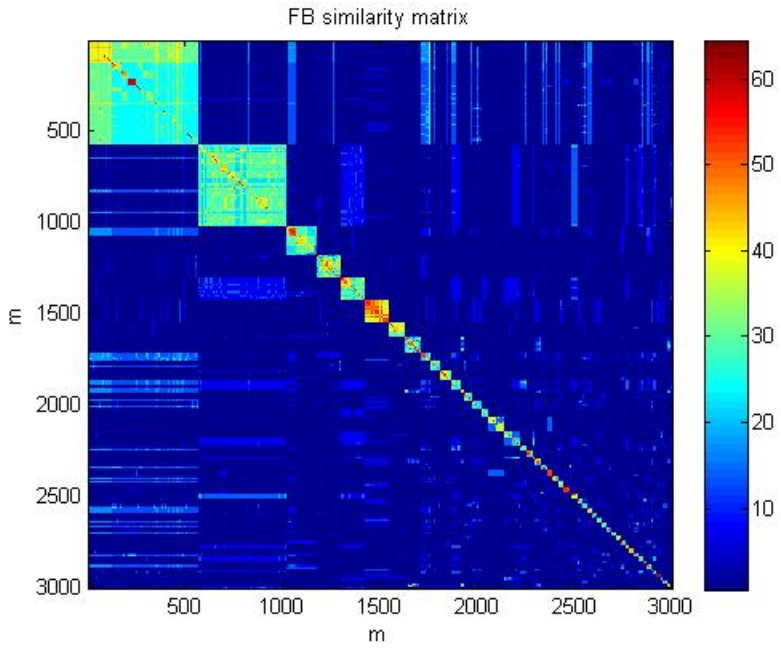

Figure 5.11: Similarity matrix permuted by FB, each entry multiplied by 60

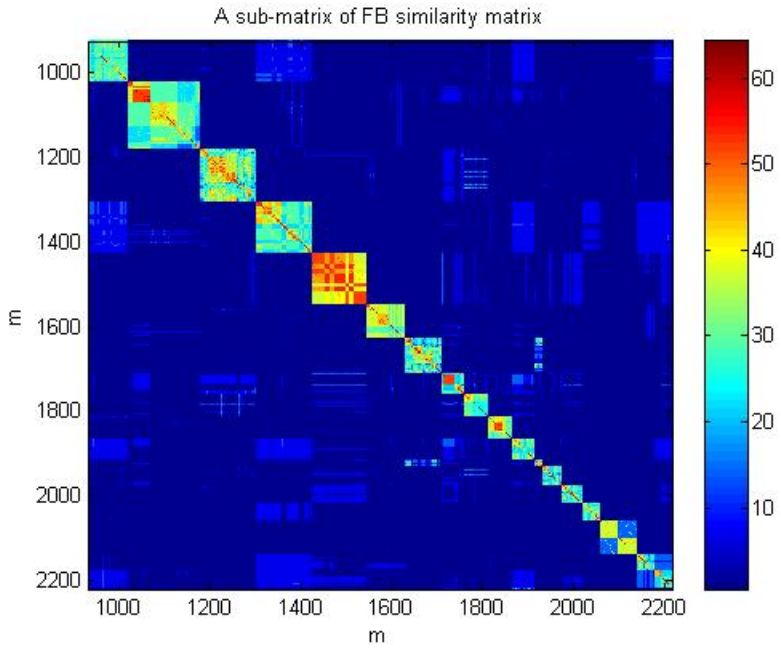

Figure 5.12: Sub-matrix of similarity matrix permuted by FB, each entry multiplied by 60

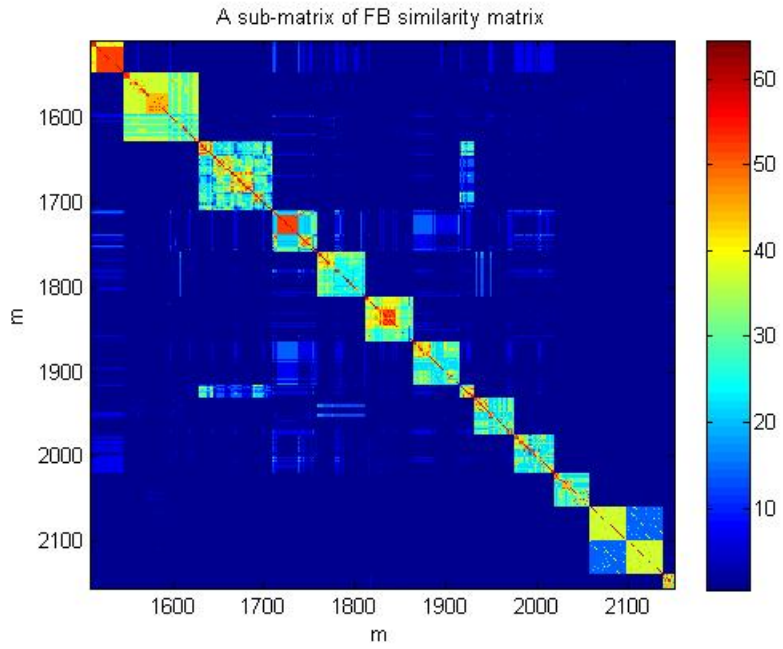

Figure 5.13: Sub-matrix of similarity matrix permuted by FB, each entry multiplied by 60

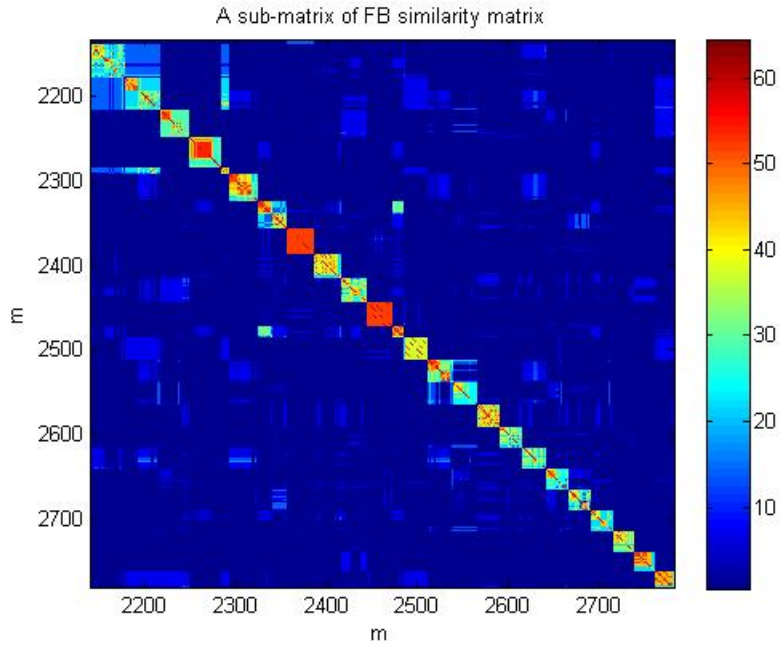

Figure 5.14: Sub-matrix of similarity matrix permuted by FB, each entry multiplied by 60

Part II

High dimensional variable selection in linear model

Chapter 6

Introduction

The linear model is widely used in many applications. Its simplest version is the following: consider $Y \in \mathbb{R}^n, \mu \in \mathbb{R}^n$, random noise $\epsilon \sim N(0, \sigma^2 I_{n\times n})$,

$$Y = \tilde{\mu} + \epsilon \tag{6.1}$$

Classically $\tilde{\mu}$ belongs to a p dimensional column space with $p \leq n$,

$$\tilde{\mu} = X\beta^* \tag{6.2}$$

with Gram matrix Σ and normalized Gram matrix Σ_n:

$$\Sigma = X^t X \ , \ \Sigma_n = \frac{X^t X}{n} \tag{6.3}$$

where $X \in \mathbb{R}^{n\times p}$, n is the number of samples and p is the number of variables. X is fixed and known, but can depend on n. Y is known but random. ϵ is unknown and random. The goal is to estimate the unknown but fixed β^*. The assumption of ϵ being i.i.d. Gaussian can be weakened in many ways. In some of the results which deal only with algorithms, as we will point out, ϵ can be an arbitrary vector. We will also point out where the assumption of i.i.d. Gaussianity or weaker assumptions, for example sub-Gaussianity, are needed.

Currently one observes many high dimensional data sets with $p > n$. In that case, β^* is unidentifiable and we need to assume further restrictions on β^*. The way out proposed by many authors, see for example, Donoho et al (D1992-1) and Chen, Donoho and Saunders (CDS1998-1), and others is to assume β^* is sparse, i.e. there is a unique β^* whose l_0 norm is upper bounded by k with $k < n$. This implies β^* is the unique solution satisfying

$$\beta^* = \arg \min_{\|\beta\|_0 \leq k} E_\epsilon[\|Y - X\beta\|_2^2] \tag{6.4}$$

Equivalently, there exists λ s.t.

$$\beta^* = \arg \min_{\beta} E_\epsilon[\|Y - X\beta\|_2^2] + \lambda\|\beta\|_0 \tag{6.5}$$

Given (X, Y) observed and under the sparsity assumption, our goal is to recover $supp(\beta^*)$ and estimate β^*.

For the classical case of $p < n$ with $rank(X) = p$, β^* could be estimated by the solution of ordinary least square(OLS):

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \|Y - X\beta\|_2^2 \tag{6.6}$$

Here $\hat{\beta}^{OLS} = (X^tX)^{-1}X^tY$. For the case of $n < p$, $\hat{\beta}^{OLS}$ is not unique. Even if $p < n$, as is well known, estimation accuracy is poor in the presence of too many predictors which are close to co-linearity, i.e. the minimal eigenvalue of the Gram matrix $\lambda_{\min}(\Sigma)$ is close to 0. An even more subtle and serious difficulty appears in the choice of which variable or group of variables are important for prediction. This difficulty is intrinsic if β^* is not unique. We will discuss this further in Chapter 10.

Under the sparsity assumption that $\|\beta^*\|_0 \leq k < n$, consider

$$\hat{\beta}^{SET} = \arg\min_{\|\beta\|_0 \leq k} \|Y - X\beta\|_2^2 \tag{6.7}$$

Suppose the following assumption holds:

Assumption 1: there exists $c > \sqrt{2}$ s.t. for $\forall L \subset [p]$ with $|L| \leq 2k$,

$$\lambda_{\min}((\Sigma_n)_{L,L}) = \lambda_{\min}(\frac{\Sigma_{L,L}}{n}) > \frac{2c\sigma}{n}\sqrt{\frac{2k\log p}{n}} \tag{6.8}$$

Then by Gaussian concentration inequality on the noise ϵ,

$$P(supp(\hat{\beta}^{set}) = supp(\beta^*)) > 1 - p^{-(c^2/2-1)} \tag{6.9}$$

Assumption 1 is a very mild condition, and we assume it holds throughout the discussion. Thus to recover β^* is equivalent to solving $\hat{\beta}^{set}$ with high probability. Unfortunately, this problem is well known to be NP-hard and exact solution is intractable. Many algorithms have been proposed to solve it approximately. Lasso (T1996-1) or equivalently Basis Pursuit (CDS1998-1) are most famous and widely used in many applications. They solve a convex relaxation of the original problem

$$\hat{\beta}_{\lambda}^{lasso} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{6.10}$$

Meinshausen and Buhlmann (MB2006-1), Zhao and Yu (ZY2006-1) and Wainwright (W2006-1) proved that the Lasso is variable selection consistent under some regularity conditions and the strong irrepresentable condition which requires $\|(\Sigma_n)_{[p]\setminus L,L}\|_\infty$ to be uniformly bounded by 1 for $\forall |L| \leq k$. Candès and Tao (CT2007-1) proposed a similar l_1 regularized Dantzig

selector $\hat{\beta}^D$ and provided an upper bound for $\|\hat{\beta}^D - \beta^*\|_2$ under the restricted isometry property(RIP) condition: $\exists \delta_k$ s.t. for $\forall |L| \leq k$ and $\forall v$,

$$(1 - \delta_k)\|v\|_2^2 \leq \frac{\|X_L v\|_2^2}{n} \leq (1 + \delta_k)\|v\|_2^2 \tag{6.11}$$

Bickel, Ritov and Tsybakov (BRT2009-1) showed that the Lasso and Dantzig selector are equivalent and provided oracle inequalities for both methods under the weakest known restricted eigenvalue(RE) condition:

$$\min_{L \subset [p], |L| \leq k} \min_{\|v_{L^c}\|_1 \leq c_0 \|v_L\|_1} \frac{\|Xv\|_2}{\sqrt{n}\|v_L\|_2} > 0 \text{ uniformly for all } (n, p, k) \tag{6.12}$$

Many other algorithms based on l_1-like regularization have been proposed. The Elastic Net by Zou and Hastie (ZH2005-1) can be considered as a general version of Lasso. It penalizes β by both l_1 and l_2 norm. The advantage is to fix unsatisfactory property of Lasso that only picks a single variable among highly correlated variables. Zou (Z2006-1) proposed adaptive Lasso to run a second stage lasso with penalization parameter adjusted by the coefficients obtained by a first stage Lasso. The advantage of this method is that the second stage Lasso removes part of the bias by penalizing less large coefficients in the first stage Lasso. A disadvantage is that if the first stage Lasso misses important variables, so does the second stage. Based on the same idea of removing bias, SCAD by J. Fan and R. Li (FL2001-1) (FL2002-1) and MC+ by C.H. Zhang (Z2010-1) are similar algorithms with non-concave penalties to penalize more confident or large coefficients less. T. Zhang proposed an iterative algorithm called adaptive forward backward selection(FoBa) (Z2011-1) and showed it behaves similarly to the Lasso.

However, real world data sets often exhibit complex covariance structures and may violate these conditions. Consider a toy counter example which is hard to detect by Lasso and other similar algorithms. Suppose

$$|cor(Y, X_1)|, |cor(Y, X_2)| >> |cor(Y, X_3)|, |cor(Y, X_4)| \tag{6.13}$$

however

$$Var(Y|X_3, X_4) << Var(Y|X_1, X_2) \tag{6.14}$$

If the solution is restricted to $\|\beta\|_0 \leq 2$, in view of the bias of $\|.\|_1$, Lasso type methods would pick $\{X_1, X_2\}$ instead of optimal $\{X_3, X_4\}$. Theoretically, this difference can be made arbitrarily large for both variable selection and prediction. We will illustrate this point by simulation in Chapter 9. Similar phenomena could potentially appear in many situations for high dimensional data. In the following chapters, we will study some such situations and provide some appropriate algorithms.

In Chapter 7, we propose a general framework to search variables based on their covariance structures. The idea is to iteratively fit small/local linear models among relatively highly correlated variables, with the fitting method potentially could be of Lasso type methods,

forward backward selection, or as simple as OLS. For simplicity, we construct the $kForward$ algorithm using OLS as the fitting step. Graphlet Screening (GS) by Jiashun Jin et al (J2012-1) and Covariance Assisted Screening and Estimation (CASE) by Tracy Ke, Jiashun Jin and Jianqing Fan (K2012-1) are similar methods which also take covariance structure into consideration, with quite different approach from ours. Their methods first screen the Gram matrix into small connected components, pick those with at least one signal variable by χ^2-test, then re-investigate each picked component with penalized MLE to remove false positives.

In Chapter 8, we analyze sufficient condition for consistent support recovery for the $kForward$ algorithm. We also show that under mild conditions, if $kForward$ initially starts with or at any step reaches the population truth $supp(\beta^*)$, then the final outcome is indeed $supp(\beta^*)$, i.e. the algorithm does not diverge from the truth once reaches it. Thus we can check if an initial procedure has identified $supp(\beta^*)$ correctly with high probability. We also propose a toy block model for the Gram matrix Σ. We show that initially start with $\emptyset$, and Σ is from the block model, $kForward$ successfully recovers $supp(\beta^*)$ under mild conditions which are strictly weaker than the RE condition.

In Chapter 9, we simulate a special case of the block model such that it violates the RE condition with extreme model parameters. For these artificially designed cases, we show by simulation that $kForward$ outperforms other methods including Lasso, Elastic Net, SCAD, MC+, FoBa. We also compare fitting and prediction performance of these algorithms in an application to US equity daily data.

In Chapter 10, we consider a different scenario where multiple mutually co-linear sets, so called minimal contexts, co-exist. Assuming an oracle algorithm exists to recover one minimal context, we construct an algorithm to systematically knock out variables from the recovered minimal contexts and call the oracle on the remaining variables. The algorithm recovers a new minimal context or guarantees there is no more such minimal context after at most k calls of the oracle, where k is the size of the minimal context. Finally we show by simulation that the algorithm works as intended.

Chapter 7

Method

7.1 General framework

Suppose $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}$ are observed,

$$Y = X\beta^* + \epsilon \tag{7.1}$$

with $\epsilon \sim N(0, \sigma^2 I_{n\times n})$. WLOG, suppose $\|\frac{X_j}{\sqrt{n}}\|_2 = 1$ and $\bar{X_j} = \sum_{i=1}^n X_{ij} = 0$ for each column X_j. Denote the Gram matrix $\Sigma = X^tX$ and normalized Gram matrix $\Sigma_n = \frac{X^tX}{n}$. Our goal is to recover $supp(\beta^*)$ and estimate β^*. Suppose Assumption 1 (6.8) holds, then it is equivalent to solve

$$\hat{\beta}^{SET} = \arg\min_{\|\beta\|_0\le k} \|Y - X\beta\|_2^2 \tag{7.2}$$

For any $J, L \subset [p] = \{1, ..., p\}$ with $|J \cup L| \le 2k$ and $s \le 2k$, define

$$\hat{\beta}^{ols}(J) = \hat{\beta}^{ols}(X, Y, J) = \arg\min_{\beta: supp(\beta)=J} \|Y - X\beta\|_2^2 \in \mathbb{R}^p \tag{7.3}$$

$$\hat{\beta}^{set}(J, s) = \arg\min_{\beta: \|\beta\|_0\le s, supp(\beta)\subset J} \|Y - X\beta\|_2^2 \in \mathbb{R}^p \tag{7.4}$$

$$\hat{\beta}^{thresh}(J, s) = (\hat{\beta}^{ols}(J)_i \mathbb{I}(|\hat{\beta}^{ols}(J)_i| \text{ is one of the top } s \text{ largest }))_{1\le i\le p} \in \mathbb{R}^p \tag{7.5}$$

$$g_J^{set}(L, s) = supp(\hat{\beta}^{set}(J \cup L, s)) \subset [p] \text{ with cardinality } s \tag{7.6}$$

$$g_J^{thresh}(L, s) = supp(\hat{\beta}^{thresh}(J \cup L, s)) \subset [p] \text{ with cardinality } s \tag{7.7}$$

where $\hat{\beta}^{ols}(J)$ is just OLS with constraint on J. $g_J^{set}(L, s)$ is the size s subset of columns of $X_{J\cup L}$ that best fit Y, and $\hat{\beta}^{set}(J \cup L, s)$ are the corresponding coefficients. $g_J^{thresh}(L, s)$ is the set of top s largest entries of $|\hat{\beta}^{ols}(J \cup L)|$, and $\hat{\beta}^{thresh}(J \cup L, s)$ are the corresponding coefficients.

Recall that under Assumption 1 (6.8), $supp(\beta^*)$ is the unique fixed point of $g_J^{set}(., k)$ for $\forall |J| \le k$, i.e. $g_J^{set}(supp(\beta^*), k) = supp(\beta^*)$. Under conditions specified in Theorem 8.1.2 later, $supp(\beta^*)$ is the unique fixed point of $g_J^{thresh}(., k)$ over all $|J| \le k$. The tradeoff

is that $g_J^{thresh}(.,k)$ is computationally much more efficient than $g_J^{set}(.,k)$. Hence if we can construct $\hat{\beta}$ s.t. $supp(\hat{\beta})$ is a fixed point of $g_J^{set}(.,k)$ or $g_J^{thresh}(.,k)$ for any $|J| \leq k$, then $supp(\hat{\beta}) = supp(\beta^*)$ with high probability. Of course this problem is not any easier than the original NP-hard problem. Our approach is to relax $\forall |J| \leq k$ to $\forall J \subset G$ with G satisfying:

1. G is reasonable to construct so that computation is feasible.
2. Under reasonable conditions, a fixed point of $g_J^{set}(.,k)$ for $\forall |J| \leq k$ is equal or close to the fixed point of $g_J^{set}(.,k)$ for $\forall J \in G$.

Our approach is to search the space of variables following their covariance structures, i.e. search variables with high covariances/correlations jointly. Consider $G = G_{s,h}$ defined as:

$$G_{s,h} = \{J_i(s,h)\}_{i=1}^p \tag{7.8}$$

$$J_i(s,h) = \{j : |\{l : |X_i^t X_j| > |X_i^t X_l|\}| \geq p - s \text{ and } |X_i^t X_j| > h\} \tag{7.9}$$

Recall that we have set $\|\frac{X_j}{\sqrt{n}}\|_2 = 1$ and $\bar{X}_j = 0$ for each column X_j. $J_i(s,h)$ is just the size s set of variables having correlation with X_i at least h (in absolute value). If there are more than s such variables, $J_i(s,h)$ consists of the top s ones having largest correlation (with X_i in absolute value). In many situations multiple $J_i(s,h)$ can overlap and coincide with each other, and $|G_{s,h}|$ can be much smaller than p. For example, if after thresholding at $O(\frac{1}{\sqrt{n}})$, Σ_n can be arranged to block diagonal with block sizes approximately equal to s, then $|G_{s,h}|$ is approximately $\frac{p}{s}$. Define the set of all size s subsets of $[p]$,

$$\Omega(s,p) = \{J \subset [p] : |J| = s\} \tag{7.10}$$

Given generic initial starting point $f_0 \subset [p]$ and generic fitting step:

$$f(.,.) : \Omega(s,p) \times (\cup_{i=0}^p \Omega(i,p)) \to \Omega(k,p) \tag{7.11}$$

our general framework to recover β^* is as follows: for $s = O(k)$,

- Standardize X, Y s.t. Y and columns of X are zero-mean and unit-variance.
- $L \leftarrow f_0$
- **while** L is not a fix point of $f(J,.)$ for all $J \in G_{s,h}$ **do**
  for each $J \in G_{s,h}$ **do**
  $L \leftarrow f(J,L)$
  end for
  end while
  return $\hat{\beta}^{ols}(L) = \arg\min_{supp(\beta)=L} \|Y - X\beta\|_2$

Under this general framework, there are many branches of algorithms based on initial starting point and specific fitting step. For example, f_0 could be support of solution of any variable selection algorithm including Lasso type of methods, or simply variables most correlated with Y, or even empty set. The generic fitting step $f(.,.)$ could be all subset regression with $f(J,L) = g_J^{set}(L,k)$, or top k largest entries of OLS with $f(J,L) = g_J^{thresh}(L,k)$, or Lasso with $f(J,L)$ being the support of size k solution on the solution path of Lasso fitted on $X_{J\cup L}$. Potentially $f(J,L)$ could be the support of size k solution of any variable selection algorithm fitted on $X_{J\cup L}$.

7.2 Algorithm

Specifically, if the fitting step $f(J,.) = g_J^{thresh}(.,k)$, i.e. the top k largest entries in absolute value of OLS on X_J, the $kForward$ algorithm is constructed as follows:

kForward(X, Y, k, s, h, M)

$(X,Y) \leftarrow$ standardize (X,Y) to be zero-mean and unit-variance
$L_0 \leftarrow$ solution of Lasso or similar methods, or largest k entries of $|Y^tX|$, or $\emptyset$
for $i = 1$ to p **do**
  $J_i(s,h) \leftarrow$ as defined in (7.9)
end for
for $iter = 1$ to M **do**
  $L_1 \leftarrow L_0$
  for $i = 1$ to p **do**
    $L \leftarrow L_0 \cup J_i(s,h)$
    $L_0 \leftarrow g_{J_i(s,h)}^{thresh}(L_0)$ or equivalently $supp(\hat{\beta}^{thresh}(L))$: the top k largest entries of $|\hat{\beta}^{ols}(L)|$
  end for
  if $L_1 == L_0$ **then**
    break
  end if
end for
return $\hat{\beta}^{kF} = \hat{\beta}^{ols}(L_0) = \arg\min_{supp(\beta)=L_0} \|Y - X\beta\|_2^2$

If the underlying model is not sparse or does not satisfy sufficient conditions stated later in Chapter 8, $kForward$ may not necessarily converge or may converge to local optima even worse than the initial starting point. If $g_J^a(k,L)$, all subset regression on $J \cup L$, is used and could be efficiently computed, then the algorithm is greedy and is guaranteed to converge to a local optimum no worse than the initial starting point. However, usually s is at least

$O(k)$ or even bigger and $g_J^a(k, L)$ is computationally expensive. Using Lasso type of methods as a fitting step could deal with larger s, but essentially would have the same convergence problem as $g_J^t(k, .)$. For better practical usage of the algorithm, the step in $kForward$:

$L_0 \leftarrow$ top k largest entries of $|\hat{\beta}|$

could be replaced by:

$L_2 \leftarrow$ top k largest entries of $|\hat{\beta}|$
if $\|Y - X\hat{\beta}(L_2)\|_2 < \|Y - X\hat{\beta}(L_0)\|_2$ **then**
$L_0 \leftarrow L_2$
end if

In other words, only update L_0 by L_2 if the later is a better OLS fit. Thus the resulting algorithm is greedy and guaranteed to converge to a local optimum at least as good as the initial starting point. For the sake of simplicity, our analysis in Chapter 8 is based on the original $kForward$ algorithm without this greedy step. However, in practice this step should be included as as to check the sufficient conditions in Chapter 8 if possible although it is NP hard.

Chapter 8

Analysis

8.1 Support Recovery

In this section we discuss sufficient conditions for the $kForward$ algorithm to recover the support of β^*. Throughout this Chapter: suppose $Y = X\beta^* + \epsilon$, $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ are known, $\epsilon\mathbb{R}^n$ is unknown, $\|X_j\|_2 = 1$ and $\bar{X}_j = \sum_{i=1}^n X_{ij} = 0$ for each column X_j, denote $\Sigma = X^tX$, $\Sigma_n = \frac{X^tX}{n}$, $L^* = supp(\beta^*)$. Suppose $\|\beta^*\|_0 = |supp(\beta^*)| = |L^*| = k < n$.

Theorem 8.1.1. *Suppose $\epsilon \in \mathbb{R}^n$ is fixed. Given L_0 and J with $|J| = s$, define:*

$$L = L_0 \cup J \tag{8.1}$$
$$U = L^*\backslash L \tag{8.2}$$
$$V = L \cap L^* \tag{8.3}$$
$$W = L\backslash L^* \tag{8.4}$$
$$\beta' = \hat{\beta}^{ols}(X, X_U\beta^*_U + \epsilon, W \cup V) \tag{8.5}$$
$$r = X_U\beta^*_U + \epsilon - X_{W\cup V}\beta' \tag{8.6}$$

where $r \perp span(X_{W\cup V})$. Recall $v_L = (v_i)_{i\in L} \in \mathbb{R}^{|L|}$, i.e.

$$\beta'_L = \beta'_{W\cup V} = \Sigma^{-1}_{W\cup V, W\cup V} X^t_{W\cup V}(X_U\beta^*_U + \epsilon) \tag{8.7}$$

Suppose $\exists s, h$ such that any $(k+s)\times(k+s)$ diagonal sub-matrix of Σ is invertible, and for $\forall |L_0| = k$, $\forall J \in G_{s,h}$, following holds:

$$\min_{i\in V} |\beta^*_i + \beta'_i| \geq \max_{j\in W} |\beta'_j| \tag{8.8}$$

Then the algorithm $kForward(X, Y, k, s, h, 1)$ successfully recovers L^*, i.e. $supp(\hat{\beta}^{kF}) = supp(\beta^*)$.

Proof. At each step of $kForward$, $L = L_0 \cup J$ for some $J \in G_{s,h}$ and $|L_0| = k$. Consider

$$\beta^0 = \hat{\beta}^{ols}(L) \tag{8.9}$$
$$\beta^0_L = \beta^0_{W\cup V} = (X^t_{W\cup V} X_{W\cup V})^{-1} X^t_{W\cup V} Y \tag{8.10}$$

Schur complement implies

$$\beta^0{}_L = \begin{pmatrix} \beta^0{}_W \\ \beta^0_V \end{pmatrix} = \begin{pmatrix} X^t_W X_W & X^t_W X_V \\ X^t_V X_W & X^t_V X_V \end{pmatrix}^{-1} \begin{pmatrix} X^t_W \\ X^t_V \end{pmatrix} (X_V \beta^*_V + X_U \beta^*_U + \epsilon) \tag{8.11}$$
$$= \begin{pmatrix} A^{-1} B (X_U \beta^*_U + \epsilon) \\ \beta^*_V + (\Sigma^{-1}_{V,V} X^t_V - \Sigma^{-1}_{V,V} \Sigma_{V,W} A^{-1} B)(X_U \beta^*_U + \epsilon) \end{pmatrix} \tag{8.12}$$

where

$$A = \Sigma_{W,W} - \Sigma_{W,V} \Sigma^{-1}_{V,V} \Sigma_{V,W} \tag{8.13}$$
$$B = X^t_W - \Sigma_{W,V} \Sigma^{-1}_{V,V} X^t_V \tag{8.14}$$

Thus by definition of β',

$$\beta^0_L = \begin{pmatrix} \beta^0_W \\ \beta^0_V \end{pmatrix} = \begin{pmatrix} \beta'_W \\ \beta^*_V + \beta'_V \end{pmatrix} \tag{8.15}$$

Hence if

$$\min_{i\in V} |\beta^*_i + \beta'_i| \geq \max_{j\in W} \beta'_j \tag{8.16}$$

V will always be selected since $kForward$ picks the top k largest entries of $|\beta^0_L|$. And since $G_{s,h} = \{J_i(s,h)\}_{i=1}^p$ contains all variables, $kForward$ recovers $supp(\beta^*)$ after searching over $G_{s,h}$ once. □

Under much milder condition than those of Theorem 8.1.1, if $kForward$ initially starts with the correct solution β^*, say obtained by some other method like Lasso, then the output is still the correct solution, i.e. $supp(\hat{\beta}^{kF}) = supp(\beta^*)$, $kForward$ will not diverge from β^*. Since the proposed various algorithms may converge to an incorrect solution, $kForward$ could be used as a check on correctness. We have,

Theorem 8.1.2. *With the same notations as in Theorem 8.1.1. Suppose $\epsilon \in \mathbb{R}^n$ i.i.d. mean-zero sub-Gaussian with parameter ρ: for $\forall t > 0, \forall \|v\|_2 = 1$,*

$$P(|v^t \epsilon| > t) < e^{-\rho t^2} \tag{8.17}$$

For example, $\epsilon \sim N(0, \sigma^2 I_{n\times n})$, $\rho = \frac{1}{2\sigma^2}$.
Furthermore, suppose $\exists s, h$, $\exists c > 1$ s.t. for $\forall J \in G_{s,h}$,

$$\lambda_{\min}(\frac{\Sigma_{J\cup L^*, J\cup L^*}}{n}) \min_{i\in L^*} |\beta^*_i| > 2c \sqrt{\frac{s\log p + k \log k}{\rho n}} \tag{8.18}$$

If we start with or at any step $L_0 = L^* = supp(\beta^*)$, then kForward recovers L^* with probability at least

$$P(supp(\hat{\beta}^{kF}) = supp(\beta^*)) \geq 1 - p^{-(c^2-1)} - k^{-(c^2-1)} \tag{8.19}$$

Remark: Notice that if $s \leq k$, $\min_{i \in L^*} |\beta_i^*| = O(1)$ and $\frac{s \log p + k \log k}{n} \to 0$, then condition (8.18) is strictly weaker than the restricted eigenvalue (RE) condition (6.12).

Proof. Recall that for $J \in G_{s,h}$, $L = J \cup L_0 = J \cup L^*$, $U = \emptyset$, $V = L^*$, $W = L \backslash L^*$ and $\beta^0 = \hat{\beta}^{ols}(L)$.

$$\beta_L^0 = \begin{pmatrix} 0 \\ \beta_{L^*}^* \end{pmatrix} + r \tag{8.20}$$

where $r = \hat{\beta}^{ols}(X, \epsilon, W \cup L^*)$ with

$$r_{W \cup L^*} = \Sigma_{W \cup L^*, W \cup L^*}^{-1} X_{W \cup L^*}^t \epsilon \tag{8.21}$$

ϵ i.i.d. sub-Gaussian implies that

$$P(|\frac{X_i^t \epsilon}{n}| > t) < e^{-n\rho t^2} \tag{8.22}$$

Let $t = c\sqrt{\frac{\log p}{\rho n}}, t' = c\sqrt{\frac{\log k}{\rho n}}$ with $c > 1$,

$$P(\exists i \text{ s.t. } |\frac{X_i^t \epsilon}{n}| > t) < pe^{-n\rho t^2} = p^{-(c^2-1)} \tag{8.23}$$

$$P(\exists i \text{ in } L^* \text{ s.t. } |\frac{X_i^t \epsilon}{n}| > t') < |L^*| e^{-n\rho t'^2} = k^{-(c^2-1)} \tag{8.24}$$

Thus with probability going to 1 for big enough c and $(p, k) \to \infty$, $X_i^t \epsilon < c\sqrt{n \log p / \rho}$ for $\forall i$, and $X_i^t \epsilon < c\sqrt{n \log k / \rho}$ for $\forall i \in L^*$,

$$\|r\|_2 \leq \lambda_{\max}(\Sigma_{W \cup L^*, W \cup L^*}^{-1}) \|X_{W \cup L^*}^t \epsilon\|_2 \tag{8.25}$$

$$\leq (\lambda_{\min}(\Sigma_{W \cup L^*, W \cup L^*}))^{-1} c\sqrt{(|W| \log p + |L^*| \log k) n / \rho} \tag{8.26}$$

$$= (\lambda_{\min}(\Sigma_{W \cup L^*, W \cup L^*}))^{-1} c\sqrt{(s \log p + k \log k) n / \rho} \tag{8.27}$$

Hence if

$$\lambda_{\min}(\frac{\Sigma_{W \cup L^*, W \cup L^*}}{n}) \min_{i \in L^*} |\beta_i^*| > 2c\sqrt{\frac{s \log p + k \log k}{\rho n}} \tag{8.28}$$

then $2 \max_j |r_j| \leq 2\|r\|_2 < \min_{i \in L^*} |\beta_i^*|$. If start with initial $L_0 = L^*$, or at any step $L_0 = L^*$, then $kForward$ selects L^* at each step afterwards. □

Corollary 8.1.3. *Suppose ϵ i.i.d. sub-Gaussian with parameter ρ as in (8.17). Suppose $L^* = supp(\beta^*)$ is successfully recovered, and we fit OLS with constraint to L^*:*

$$\hat{\beta}^{ols}(L^*) = \arg \min_{supp(\beta)=L^*} \|Y - X\beta\|_2^2 = \Sigma_{L^*,L^*}^{-1} X_{L^*} Y \tag{8.29}$$

Then for $\forall c > 1$

$$P(\|\hat{\beta}^{ols}(L^*) - \beta^*\|_2 \leq c(\lambda_{\min}(\frac{\Sigma_{L^*,L^*}}{n}))^{-1} \sqrt{\frac{k \log k}{\rho n}}) \geq 1 - k^{-(c^2-1)} \tag{8.30}$$

Proof. Use the upper bound of $\|r\|_2$ given by (8.25) – (8.27). □

8.2 Special Block Model

From Theorem 8.1.1, it can be deduced that $kForward$ recovers $supp(\beta^*)$ if $\Sigma = X^t X$ is of some special forms. One case would be that Σ is partially block diagonal, i.e. $supp(L^*)$ is a subset of union of small blocks of Σ. Specifically, consider $\Sigma \in F_{h,s}(L^*)$,

$$F_{h,s}(L^*) = \{\Sigma : \text{ for } \forall i \in L^*, \exists B_i \subset [p] \text{ with } i \in B_i \text{ and } |B_i| \leq s \tag{8.31}$$

$$\text{s.t. } |\Sigma_{jl}| > h \text{ if and only if } (j,l) \in B_i \times B_i\} \tag{8.32}$$

This is to require each important variable $i \in L^*$ is contained in a block B_i with size at most s, such that variables within B_i are highly correlated with correlation at least h, while correlation between variables inside and outside B_i are small and upper bounded by h. This is a strong sufficient requirement, while the algorithm actually works for much more general situations as shown in the simulation and application in Chapter 9. Also notice that $F_{h,s}(L^*)$ is similar to the class of sparsifiable Gram matrices proposed in CASE by Ke et al (K2012-1). Next theorem shows that if $\Sigma \in F_{h,s}(L^*)$, $kForward$ fully recovers $supp(\beta^*)$ under mild conditions.

Theorem 8.2.1. *With the same notations as in Theorem 8.1.2. Suppose ϵ i.i.d. sub-Gaussian with parameter ρ as in (8.17). Suppose the Gram matrix $\Sigma \in F_{h,s}(L^*)$, $\exists c > 1$ s.t. for $\forall L \subset [p]$ with $|L| \leq s + k$,*

$$\lambda_{\min}(\frac{\Sigma_{L,L}}{n}) \min_{i \in L^*} |\beta_i^*| > 2c \sqrt{\frac{s \log p + k \log k}{\rho n}} + \frac{h}{n} \sqrt{k(s+k)} \|\beta^*\|_2 \tag{8.33}$$

Then if we start with $L_0 = \emptyset$, kForward$(X, Y, k, s, h, 1)$ recovers L^* with probability at least

$$P(supp(\hat{\beta}^{kF}) = supp(\beta^*)) \geq 1 - p^{-(c^2-1)} - k^{-(c^2-1)} \tag{8.34}$$

Proof. Suppose $kForward$ starts with $L_0 = \emptyset$. At any step, for $J \in G_{s,h}$, if $L = L_0 \cup J, U = L^* \backslash L, V = L \cup L^*, W = L \backslash L^*, \beta^0 = \hat{\beta}^{ols}(L)$, Theorem 8.1.1 implies

$$\beta^0_L = \beta^0_{W \cup V} = \begin{pmatrix} 0 \\ \beta^*_V \end{pmatrix} + r^U + r^\epsilon \tag{8.35}$$

where $r^U = \hat{\beta}^{ols}(X, X_U \beta^*_U, W \cup V)$ and $r^\epsilon = \hat{\beta}^{ols}(X, \epsilon, W \cup V)$,

$$r^U_{W \cup V} = \Sigma^{-1}_{W \cup V, W \cup V} X^t_{W \cup V} X_U \beta^*_U \tag{8.36}$$

$$r^\epsilon_{W \cup V} = \Sigma^{-1}_{W \cup V, W \cup V} X^t_{W \cup V} \epsilon \tag{8.37}$$

$\Sigma \in F_{h,s}(L^*)$ and the construction of $F_{h,s}(L^*)$ imply

$$|\Sigma_{jl}| \leq h \text{ for } \forall (j,l) \in (W \cup V) \times U \tag{8.38}$$

Hence

$$\|r^U\|_2 \leq \lambda_{\max}(\Sigma^{-1}_{W \cup V, W \cup V}) \|\Sigma_{W \cup V, U} \beta^*_U\|_2 \tag{8.39}$$

$$\leq (\lambda_{\min}(\Sigma_{W \cup V, W \cup V}))^{-1} h \sqrt{|U||W \cup V|} \|\beta^*_U\|_2 \tag{8.40}$$

$$\leq (\lambda_{\min}(\Sigma_{W \cup V, W \cup V}))^{-1} h \sqrt{k(s+k)} \|\beta^*\|_2 \tag{8.41}$$

Similar to Theorem 8.1.2, for $\forall c > 1$, with probability at least $1 - p^{-(c^2-1)} - k^{-(c^2-1)}$,

$$\|r^\epsilon\|_2 \leq (\lambda_{\min}(\Sigma_{W \cup V, W \cup V}))^{-1} c \sqrt{(s \log p + k \log k) n / \rho} \tag{8.42}$$

Hence if $\exists c > 1$ s.t.

$$\lambda_{\min}(\frac{\Sigma_{W \cup V, W \cup V}}{n}) \min_{j \in L^*} |\beta^*_j| > 2c \sqrt{\frac{s \log p + k \log k}{\rho n}} + \frac{h}{n} \sqrt{k(s+k)} \|\beta^*\|_2 \tag{8.43}$$

then

$$2 \max_j |(r^U + r^\epsilon)_j| \leq 2 \|r^U + r^\epsilon\|_2 < \min_{j \in V} |\beta^*_j|$$

This implies $kForward$ always includes V. Since $G_{s,h} = \{J_i(s,h)\}_{i=1}^p$ contains all variables, $kForward$ recovers $supp(\beta^*)$ after searching over $G_{s,h}$ once. □

Chapter 9

Experiment

9.1 Simulation

In this section we compare support recovery and prediction performances of $kForward$ to other popular algorithms including Lasso(T1996-1), Elastic Net (ZH2005-1), SCAD (FL2001-1), FoBa (Z2011-1), MC+ (Z2010-1). We construct a special case of $F_{h,s}(L^*)$ such that, for extreme model parameters, it violates the RE condition: L^* belong to $m = \frac{k}{s}$ blocks of size s, but the minimal eigenvalue of each block goes to 0. This example is unrealistic, but illustrates theoretically what extreme cases could lead to. This type of example was suggested to us by Boaz Nadler and Ya'acov Ritov.

Specifically, for $w \in \mathbb{R}^m, \epsilon \sim N(0, \sigma^2 I_{n\times n})$, let

$$Z = (Z_1, Z_2, ..., Z_{ms}) \in \mathbb{R}^{n\times ms} \text{ i.i.d. } N(0, I_{n\times n}) \tag{9.1}$$

$$\mathcal{Z}_j = \sum_{i=(j-1)s+1}^{js} Z_i \text{ for } j = 1, ..., m \tag{9.2}$$

$$\mathbb{X}_i = w_j(Z_i - \frac{Z_i^t \mathcal{Z}_j}{\|\mathcal{Z}_j\|_2^2}\mathcal{Z}_j) + \mathcal{Z}_j \text{ for } i = (j-1)s+1, ..., js,\ j = 1, ..., m \tag{9.3}$$

$$X_i = \frac{\mathbb{X}_i}{\sqrt{\mathbb{X}_i^t \mathbb{X}_i / n}} \text{ for } i = 1, ..., ms \tag{9.4}$$

$$X_i \sim N(0, I_{n\times n}) \text{ for } i = ms+1, ..., p \tag{9.5}$$

$$Y = s\sum_{j=1}^{m} \mathcal{Z}_j + \epsilon \tag{9.6}$$

where w is a parameter controlling difficulty of support recovery. Specifically, for $p = 1000, n = 100, k = 9, m = s = 3, \sigma = 1$, i.e. $supp(\beta^*)$ is contained in 3 small blocks of size 3. By a simulation of 1000 iterations, some empirical statistics for relevant parameters are shown in Table 9.1. Some immediate observations:

w	$\lvert cor(X_1, Y)\rvert$	$\lvert cor(X_1, X_2)\rvert$	$P(\Sigma \in F_{h,k})$	$\lambda_{\min}(\frac{\Sigma_{[s],[s]}}{n})$	$\lambda_{\min}(\frac{\Sigma_{[s],[s]}}{n}) \min_{i\in[k]} \lvert\beta_i^*\rvert$
1	0.51±0.06	0.72±0.09	1	0.21±0.05	0.38±0.09
2	0.42±0.05	0.29±0.16	0.23	0.56±0.11	1.27±0.25
3	0.33±0.05	0.07±0.11	0	0.79±0.12	2.28±0.37
4	0.27±0.04	0.17±0.17	0	0.59±0.13	2.10±0.45
5	0.22±0.03	0.27±0.16	0.02	0.41±0.09	1.71±0.37
7	0.17±0.03	0.37±0.15	0.50	0.22±0.05	1.27±0.30
10	0.12±0.02	0.43±0.15	0.87	0.11±0.03	0.91±0.22
1000	0.07±0.04	0.49±0.15	0.97	1.2e-5±3.1e-6	0.009±0.002

Table 9.1: $p = 1000, n = 100, k = 9, m = s = 3, \sigma = 1$

- $\lvert cor(X_1, Y)\rvert$ converges to 0 as w grows. This suggests the signal from a single variable goes to 0. In some sense, the difficulty of the problem grows with w.
- $\lambda_{\min}(\frac{\Sigma_{[s],[s]}}{n})$ converges to 0 as w grows. Hence for moderately small w, the RE condition is satisfied and Lasso type of methods are expected to work well.
- For big w, $P(\Sigma \in F_{h,k})$ is converging to 1. This is the part of sufficient conditions for $kForward$ to work well.
- For small w, $\lambda_{\min}(\frac{\Sigma_{[s],[s]}}{n}) \min_{i\in[k]} \lvert\beta_i^*\rvert$ are relatively big. This is also a part of sufficient conditions for $kForward$ to work well. However, it converges to 0 as w grows, which violates the condition for big w.

(X^π, Y) are observed, where the columns $X_i^\pi = X_{\pi(i)}$ with π being a random permutation of $[p]$. Following algorithms are compared:

1 Lasso: Pick solution containing k variables from the solution path.

2 elastic net: Use cross validation to choose the weight between l_1 and l_2 penalties. For each given weight, pick solution with k variables from the solution path.

3 FoBa: Use cross validation to choose algorithm parameter. For given parameter, pick solution with k variables from the solution path.

4 SCAD: Use cross validation to choose algorithm parameter. For given parameter, pick solution with k variables from the solution path.

5 MC+: Use cross validation to choose algorithm parameter. For given parameter, pick solution with k variables from the solution path.

6 kF^c: Run $kForward$ algorithm with $G_{k,0}$ and initially start with k variables that most correlated with Y.

7 kF^l: Run $kForward$ algorithm with $G_{k,0}$ and initially start with k variables picked by Lasso.

8 kF^o: Between the solutions of kF^c and kF^l, choose the one with better training R^2.

Cases $p = 1000, 2000, 3000$ are simulated, with $n = 100, k = 9, m = s = 3, \sigma = 1$ fixed. Both training set and test set consists of 100 samples. Run each algorithm on the training set to recover k variables, say $|\hat{\beta}| = k$. The performance is evaluated by two criteria: the Hamming distance and the test R^2. The Hamming distance between $supp(\beta^*)$ and $supp(\hat{\beta})$ is defined as $|supp(\beta^*) \backslash supp(\hat{\beta})|$. The test R^2 is obtained by applying training OLS fit on $supp(\hat{\beta})$ to the test set. The test R^2 is used because in many experiments in previous literature, differences in variable selection are well shown, but not simultaneously about prediction. We want to show that in this artificial case, there are huge differences for both variable selection and prediction. The result for Hamming distance is reported in Figure 9.1, and the result for test R^2 is reported in Figure 9.2. Instead of Gaussian, the case of X, Z, ϵ following Laplace distribution is also simulated. The corresponding results are reported in Figure 9.3 and Figure 9.4. Notice that the results are very similar to each other.

9.2 Application

Although not as extreme as the results of previous artificial example, in this section we show that there are indeed significant differences among algorithms for a real world example. As there is no correct answer but only sparse approximation for the real world data, the only comparison criteria is fitted R^2.

Consider US equity daily data consisting of 2316 stocks and 47 ETFs from 2007 January 1st to 2012 December 31st. For each day and for each ETF, 6 algorithms as in previous section, i.e. Lasso (T1996-1), Elastic Net(ZH2005-1), FoBa (Z2011-1), SCAD (FL2001-1), kF^c and kF^l, are used to select k stocks using 100 samples consisting of past 50 days' open and close prices. Then OLS is fitted on the picked k stocks and corresponding R^2 is calculated. For $k = 5$, average R^2 over 6 years are reported in Table 9.2. Since there are stocks having correlations with ETFs higher than 0.98, each algorithm works well and pretty much similarly for k as small as 5. We increase the difficulty of the problem by requiring that only weak signals are allowed in the model, i.e. only stocks having correlation with the response ETF lower than a certain threshold are included in the model. In an experiment, correlation threshold $0.1, 0.2, 0.3, 0.4$ and $k = 5, 10, 15$ are used. The results are reported in Table 9.3, Table 9.4, Table 9.5, Figure 9.5 and Figure 9.6. As shown by the result, the improvement of $kForward$ is most clear for the most difficult cases, i.e. k is small and correlation threshold is small.

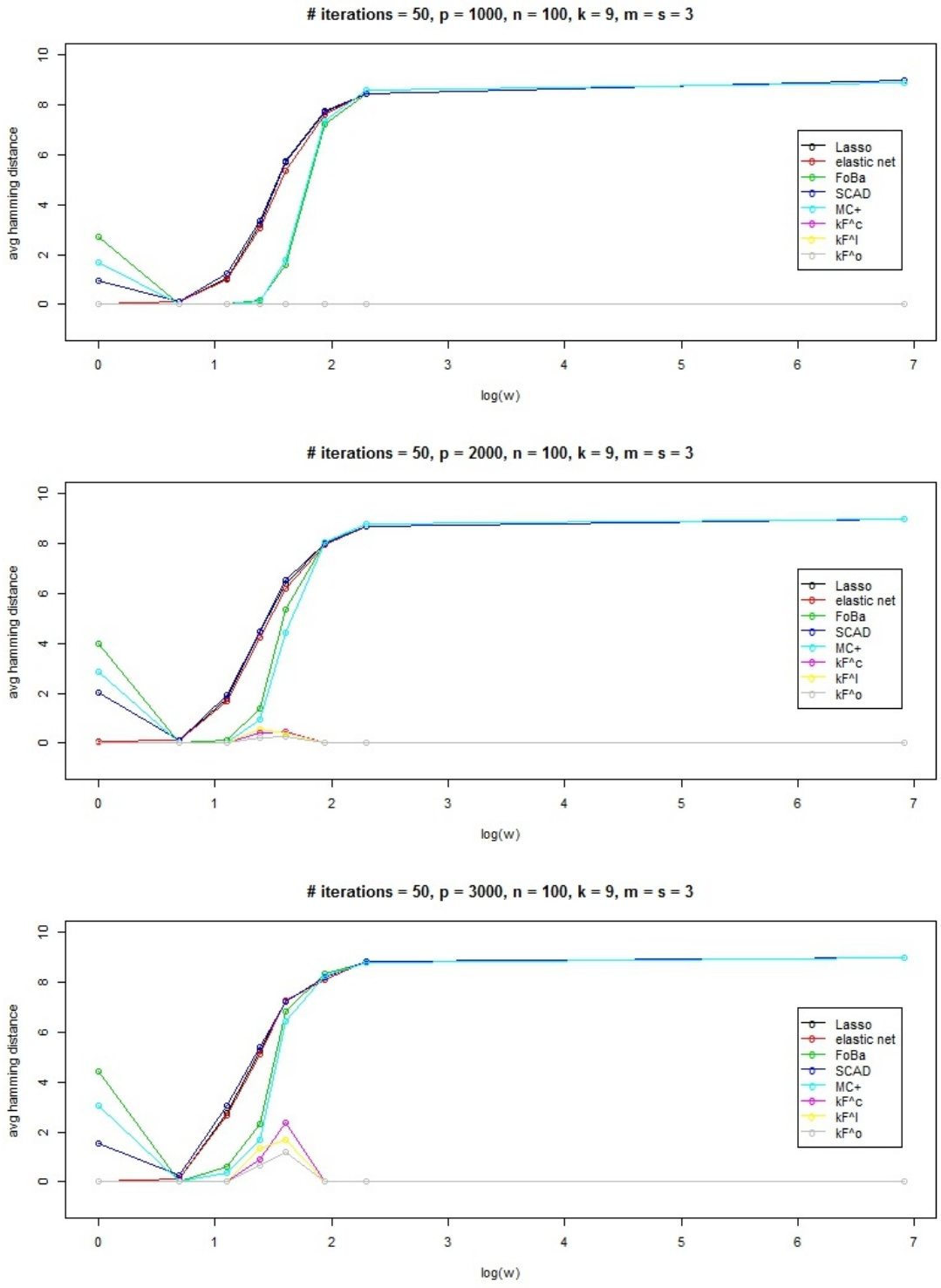

Figure 9.1: $\log(w)$ v.s. average Hamming distance between $supp(\beta^*)$ and $supp(\hat{\beta})$. X, Z, ϵ follows Gaussian distribution

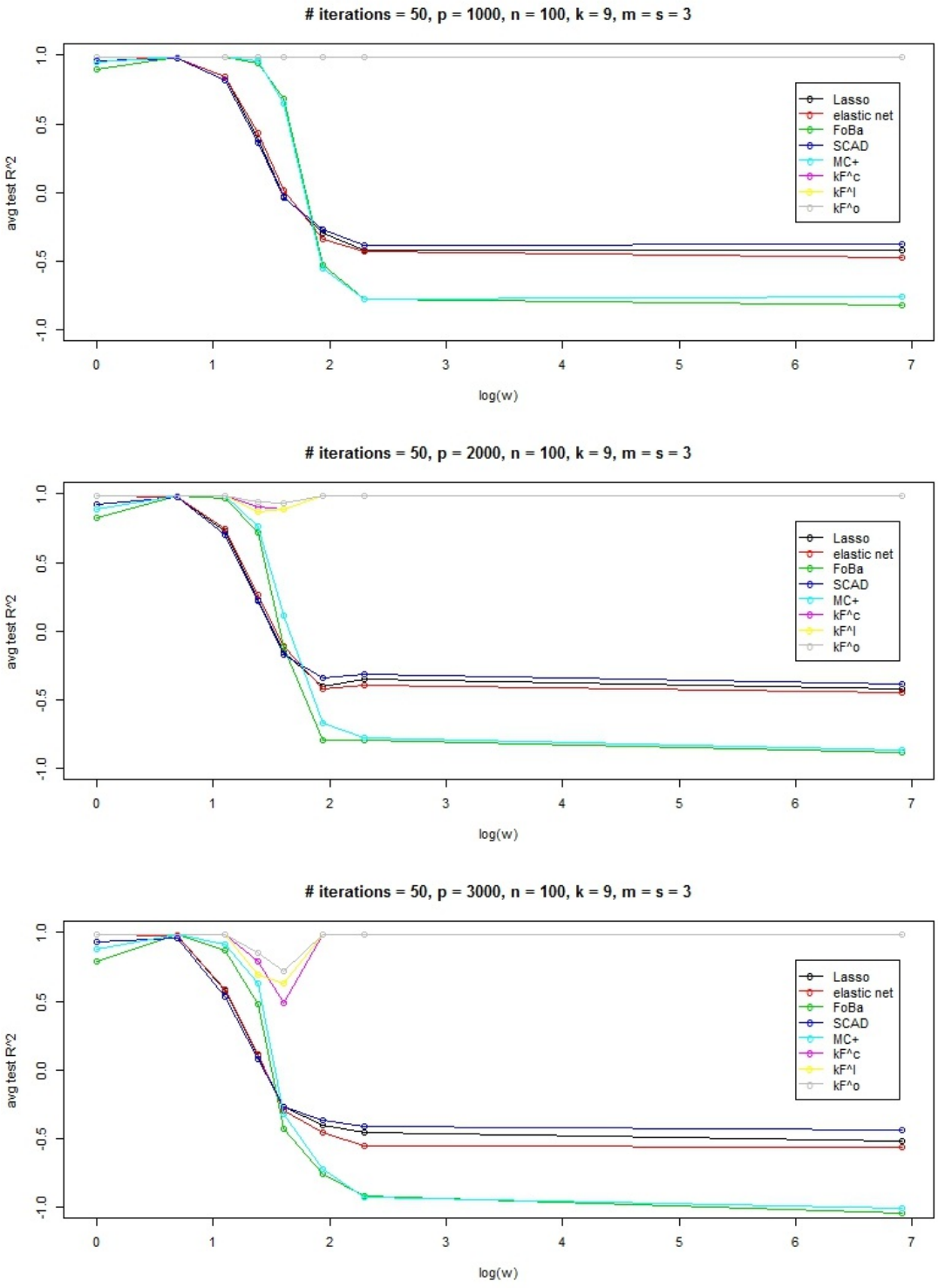

Figure 9.2: $\log(w)$ v.s. average test R^2. X, Z, ϵ follows Gaussian distribution

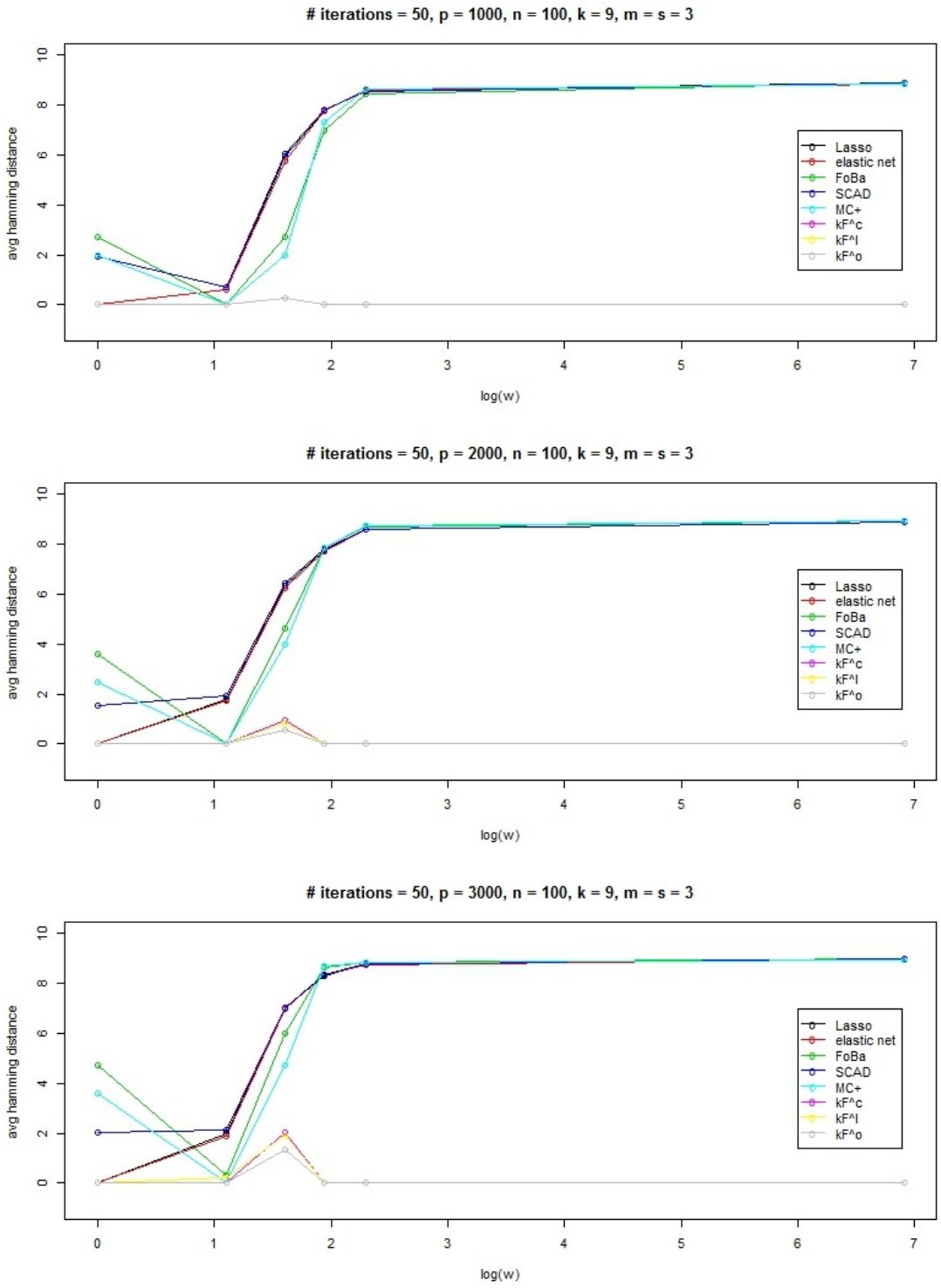

Figure 9.3: $\log(w)$ v.s. average Hamming distance between $supp(\beta^*)$ and $supp(\hat{\beta})$. X, Z, ϵ follows Laplace distribution

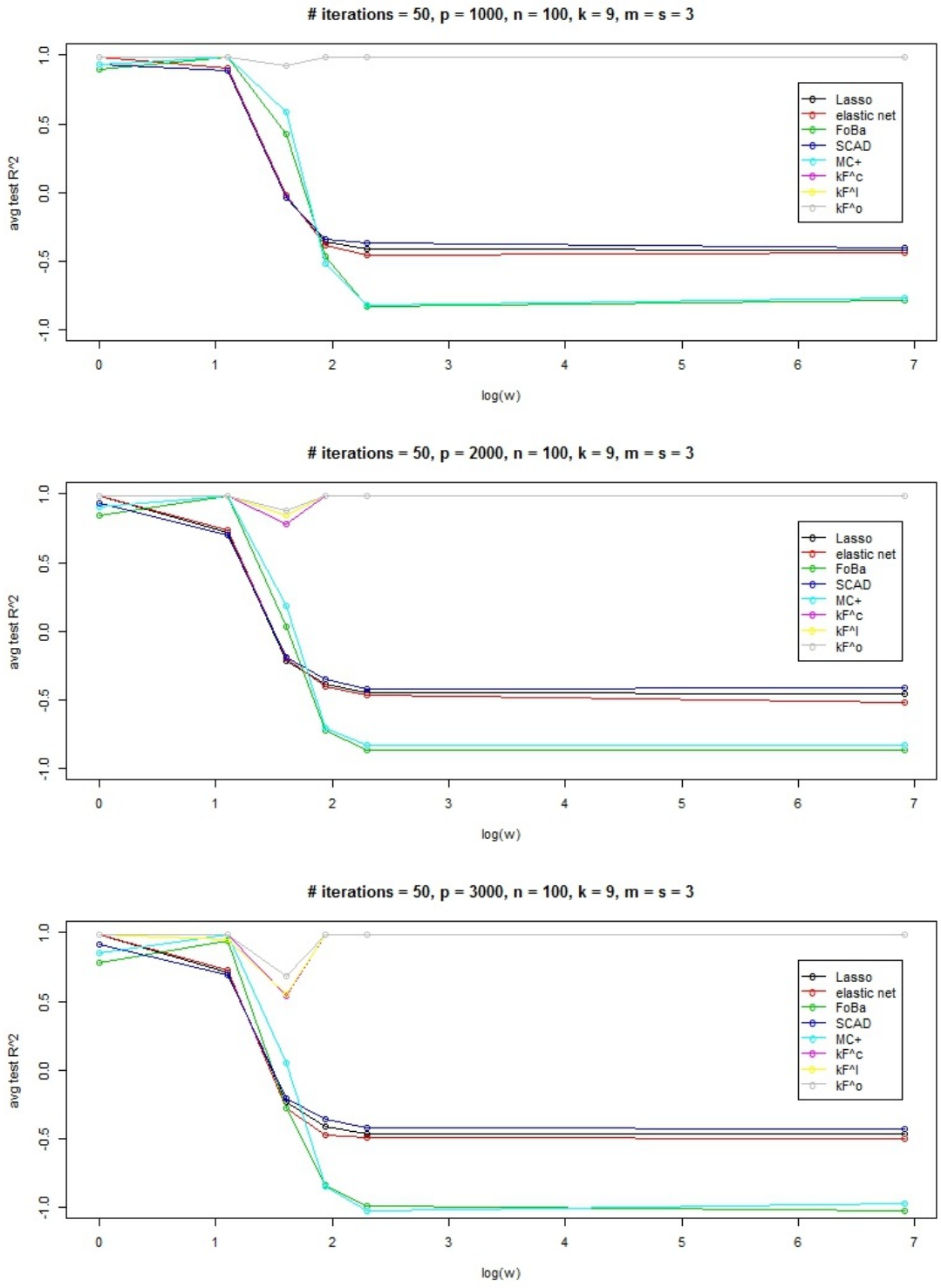

Figure 9.4: $\log(w)$ v.s. average test R^2. X, Z, ϵ follows Laplace distribution

	Lasso	elastic net	FoBa	SCAD	kF^c	kF^l
R^2	0.968	0.972	0.973	0.967	0.986	986

Table 9.2: average R^2 for $k = 5$, no correlation threshold

correlation threshold	Lasso	elastic net	FoBa	SCAD	kF^c	kF^l
0.1	0.30	0.36	0.30	0.21	0.49	0.50
0.2	0.62	0.65	0.62	0.47	0.79	0.80
0.3	0.76	0.80	0.81	0.63	0.90	0.89
0.4	0.82	0.86	0.87	0.72	0.93	0.93

Table 9.3: average R^2 for $k = 5$

correlation threshold	Lasso	elastic net	FoBa	SCAD	kF^c	kF^l
0.1	0.52	0.59	0.57	0.50	0.74	0.75
0.2	0.79	0.82	0.85	0.76	0.93	0.93
0.3	0.88	0.90	0.93	0.87	0.96	0.96
0.4	0.91	0.93	0.95	0.90	0.98	0.98

Table 9.4: average R^2 for $k = 10$

correlation threshold	Lasso	elastic net	FoBa	SCAD	kF^c	kF^l
0.1	0.69	0.73	0.73	0.70	0.85	0.85
0.2	0.88	0.88	0.92	0.88	0.97	0.96
0.3	0.94	0.94	0.96	0.94	0.98	0.98
0.4	0.95	0.95	0.97	0.95	0.99	0.99

Table 9.5: average R^2 for $k = 15$

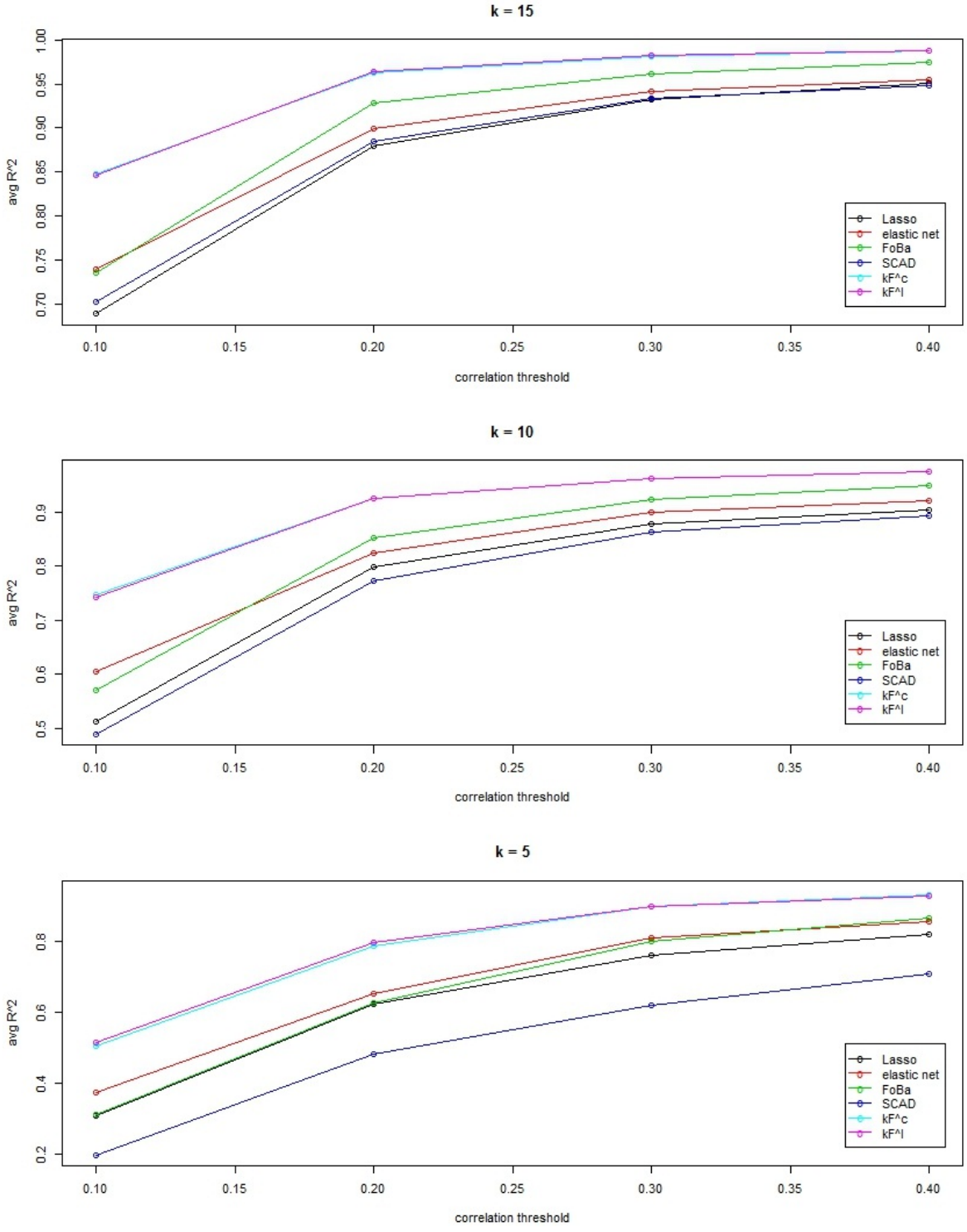

Figure 9.5: correlation threshold v.s. average R^2 for $k = 5, 10, 15$

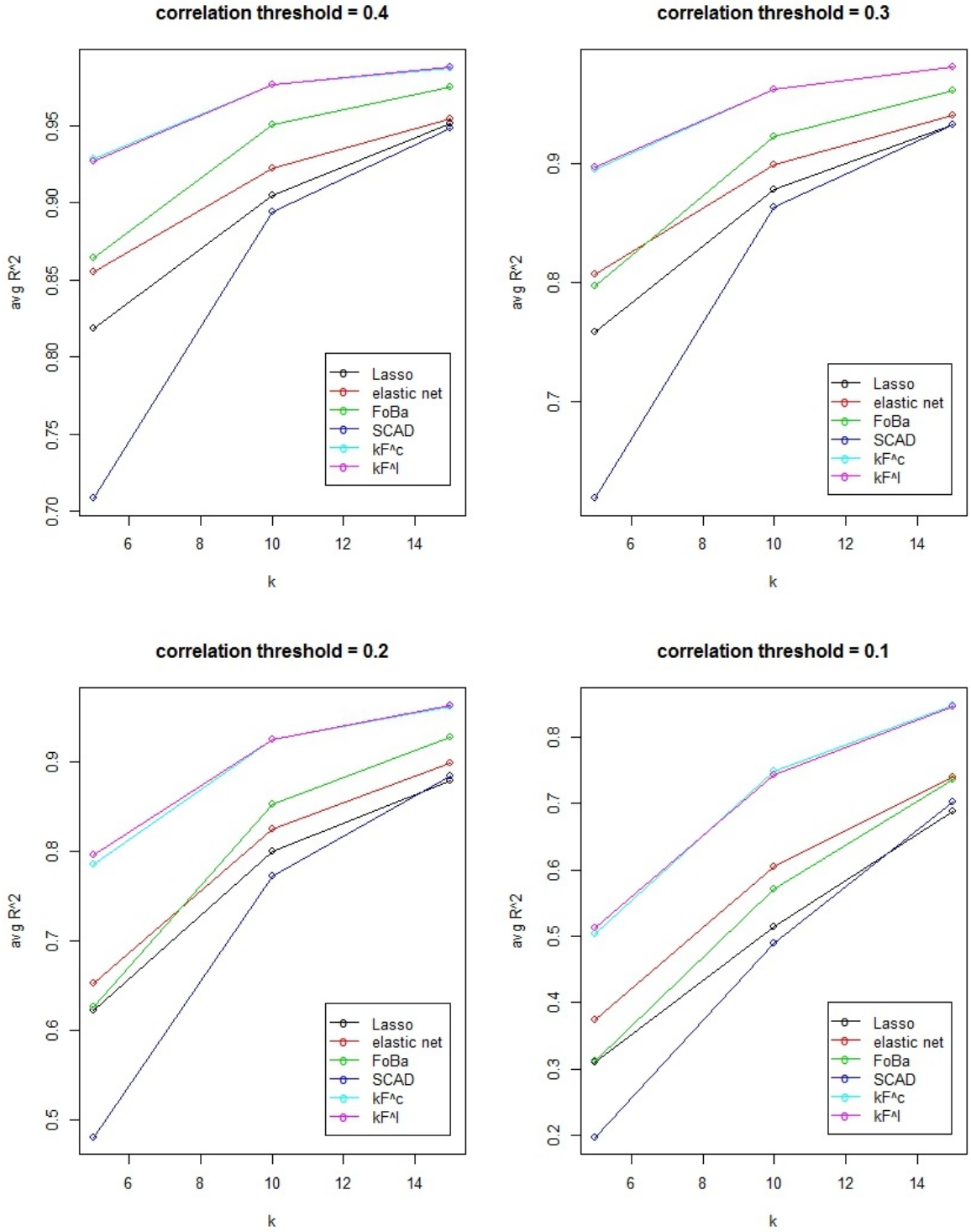

Figure 9.6: k v.s. average R^2 for correlation threshold $= 0.1, 0.2, 0.3, 0.4$

Chapter 10

Minimal Context

10.1 A Different Framework[1]

Suppose $X = (\mathbf{Z}, Y)$, $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$, and we have a sample $(\mathbf{Z}_i, Y_i)$, $i = 1, \ldots, n$, and $\mu(\mathbf{Z}) \equiv E(Y|\mathbf{Z})$ unknown. We typically approximate μ by

$$\mu_N(\mathbf{Z}) \equiv \sum_{j=1}^{N} \beta_{jN} g_j(\mathbf{Z}) + \beta_{0N}$$

where $g_0 \equiv 1, g_1, \ldots, g_N, \ldots$ is a basis for $L_2(\mathbf{Z})$ and

$$\mu_N(\mathbf{Z}) \equiv \arg\min \|Y - \sum_{j=0}^{N} \beta_j g_j(Z)\|^2$$

For simplicity, we take $\mu_N \equiv \mu$ with the understanding that N changes with n. The selection of variables question informally is, "Which factors Z_j are important?" The g_j are, typically, functions of several variables bringing in a major complication. We consider only the simple case $g_j(\mathbf{Z}) = Z_j$, $N = p$ which illustrates the issues we raise. It is reasonable to measure effectiveness of a set S of factors by its predictive power in relation to optimal prediction. Formally, we define,

$$\begin{aligned} r(S) &\equiv 1 - \Big\{ \frac{\|Y - \mu(\mathbf{Z}, S)\|^2 - \|Y - \mu(\mathbf{Z})\|^2}{\mathrm{Var}(Y)} \Big\} \\ &= \frac{\|\mu(\mathbf{Z}) - \mu(\mathbf{Z}, S)\|^2}{\sigma^2 + \|\mu(\mathbf{Z})\|^2} \end{aligned}$$

[1] **Chapter 10.1 and 10.3 are from "Discussion of Sara van de Geer: generic chaining and the L_1 penalty", Peter Bickel and Mu Cai, submitted to Journal of Statistical Planning and Inference.**

where $\sigma^2 \equiv \|Y - \mu(\mathbf{Z})\|^2$, and

$$\mu(\mathbf{Z}, S) = \beta_0 + \sum\{\beta_j(S)Z_j : Z_j \in S\}$$

where β_0 is the intercept, and $\beta(S)$ are determined by

$$\mu(\mathbf{Z}, S) = \arg\min \|Y - \beta_0 - \sum_{j=1}^{p} \beta_j Z_j : Z_j \in S\}\|^2$$

Remarks: 1) $\beta_j(S)$ depend on S unless $Z_j \perp Z_k$ all $j \neq k$. From this point of view (B2011-1) the importance of a factor depends on its *context* the other variables that are in the set S of Z's being considered as predictors.

2) It is entirely possible in the case of collinearity to have $r(S_1) = r(S_2) = 1$, $S_1 \neq S_2$ and, in general, to have a class of sets $\mathcal{S}_{\varepsilon,m}$ with $r(S) \geq 1 - \varepsilon$ for $S \in \mathcal{S}_{\varepsilon,m}$ and $|S| \leq m$ for $m(\varepsilon)$ sufficiently large

We define the relative *contribution* of Z_j to the predictive power of S by

$$\text{(a)} \qquad c(Z_j, S) = \frac{\|\mu(\mathbf{Z}, S_{-j}) - \mu(\mathbf{Z}, S)\|^2}{\|\mu(\mathbf{Z})\|^2}$$

where $S_{-j} = \{Z_k : k \in S,\ k \neq j\}$.

This leads us to the following approach to the importance of a variable Z_j in a context S.

(i) We want the context to have high predictive power, $r(S) \geq 1 - \varepsilon$

(ii) We want the context as small as possible for interpretability, $|S| \leq s_0$

(iii) We want the contribution of Z_j, $c(Z_j, S)$ to the predictive power of the context be high.

Next we note that by orthogonality,

$$\begin{aligned} \text{(b)} \qquad c(Z_j, S) &= \frac{\beta_j^2(S)}{\|\mu(\mathbf{Z})\|^2} E\big(Z_j - \Pi(Z_j|[S_{-j}])\big)^2 \\ &= (Y, Z_j - \hat{Z}_j)^2 / \|\mu(\mathbf{Z})\|^2 \|Z_j - \hat{Z}_j\|^2 \end{aligned}$$

where $\Pi(\cdot|L)$ is L_2 projection onto a linear space, $[S] \equiv$ Linear span of S, and

$$\hat{Z}_j \equiv \Pi(Z_j|[S_{-j}]) \ .$$

We can also write

$$\text{c)} \qquad c(Z_j, S) = \big(1 + \frac{\sigma^2}{\|\mu(\mathbf{Z})\|^2}\big)\ \text{corr}^2\big(Y, Z_j - \hat{Z}_j\big) \ .$$

We introduce two more concepts. S is $\boldsymbol{\epsilon}$ minimal if $r(S) \geq 1-\varepsilon$ and $r(S_{-j}) \leq 1-2\varepsilon$ for all $j \in S$.

Finally Z_j is $\boldsymbol{\delta}$ important in minimal ε context S iff $c(Z_j, S) \geq 1-\delta$.

Our goal is, having chosen δ, ε to find all (δ, ε) relevant factors, as above.

If p is small we know how to solve the problem using all subsets regression. What if p is large? We assume

A0: There is a sparse representation, i.e. For S, $|S| \leq m_0 < \infty$ independent of p, n

$$E\big(Y - \mathbf{Z}^T(S)\boldsymbol{\beta}(S)\big)^2 = \arg\min E(Y - \mathbf{Z}^T\boldsymbol{\beta})^2 \ .$$

A1: For any ε minimal context, $|S| \leq s_0(\varepsilon) \leq m_0$.

A2: The set of all ε minimal contexts, $\mathcal{C}(\varepsilon)$, has

$$|\mathcal{C}(\varepsilon)| \leq M_0$$

A3: The minimal eigenvalues of all Gram matrices of minimal contexts is $\geq \tau > 0$.

A4: The distribution of Y is subGaussian, for all $t > 0$

$$Ee^{tY} \leq \exp\{\frac{t^2\lambda^2}{2}\}$$

where $\lambda^2 \equiv \mathrm{Var}(Y)$

A5: $|Z_j| \leq M$ all $j = 1, \ldots, p$.

Essentially we are ruling out situations where good prediction is achieved by combining a large number of factors each contributing negligibly – not because we do not believe such situations exist but because we cannot usefully distinguish what factors are important in such cases.

Proposition 10.1.1. *If we ignore computational considerations then under A1-A5, even if $p, n \to \infty$ we can identify all ε minimal contexts and δ important factors within them if $\frac{\log p}{n} \to 0$.*

Proof. Since by (A1) the size of all ε minimum contexts is bounded by s_0, it suffices to show that we can find a consistent estimate $\hat{t}$ of the minimal prediction error t and then, among all subset of factors of cardinality $\leq m_0$, find those sets with empirical predictive error $\geq (1-\varepsilon)\hat{t}$ and then among those, identify the δ important factors. Suppose we have $\hat{t}$. We claim that it is then enough to show that for any S, $|S| \leq s_0$,

$$P\big[\hat{\boldsymbol{\beta}}(S) - \boldsymbol{\beta}(S) \geq \varepsilon\big] \leq Ce^{-\varepsilon^2 n\gamma} \tag{10.1}$$

for c, γ independent of s_0. We can then examine all $\binom{p}{m}$, $m \leq s_0$ regressions of Y on m factors, $m \leq s_0$. The union sum inequality applied to (1) and the condition of the proposition ensures that the minimum empirical MSE of regressions based on $\leq m_0$ factors converges to the population MSE. We need only slightly refine the results of Fu and Knight (FK2000-1) for p fixed, $E\mathbf{Z} = \mathbf{0}$. Write

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i\mathbf{Z}_i^T(S) - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T(S)$$

where $\mathbf{Z}_i(S) \equiv \{Z_{ij} : \ j \in S\}_{s_0 \times 1}$.

$$\Sigma(S) = E\mathbf{Z}_1\mathbf{Z}_1^T(S) \ .$$

By definition,

$$\begin{aligned} \frac{1}{n}[\tilde{Z}_S]\mathbf{Y} &= \frac{1}{n}[\tilde{Z}_S][\tilde{Z}_S]^T\hat{\boldsymbol{\beta}}(S) \\ E\mathbf{Z}_1(S)Y &= \Sigma(S)\boldsymbol{\beta}(S) \\ \mathbf{Y} &\equiv (Y_1, \ldots, Y_n)^T \\ [Z_S] &\equiv [\mathbf{Z}_1(S), \ldots, \mathbf{Z}_n(S)]_{s_0 \times n} \\ [\tilde{Z}_S] &\equiv [Z_S] - \big(\bar{\mathbf{Z}}(S), \ldots, \bar{\mathbf{Z}}(S)\big) \end{aligned}$$

By assumptions A4, A5

$$P\big[\big|\frac{[\tilde{Z}_S]Y}{n} - E\tilde{\mathbf{Z}}(S)Y\big| \geq t\big] \leq 2e^{-\frac{t^2 n}{2M^2\lambda^2}} \tag{10.2}$$

since

$$Ee^{t(Z_{ij}Y_i - EZ_{ij}Y_i)} \leq e^{t^2M^2\lambda^2/2}$$

where $\lambda^2 = \mathrm{Var}Y$.

Also,

$$\frac{1}{n}[\tilde{Z}_S][\tilde{Z}_S]^T = \hat{\Sigma}(S) \ .$$

We can apply Oliveira's (O2010-1) inequality to obtain

$$P\big[\frac{1}{n}\|\sum_{i=1}^{n}\big(\mathbf{Z}_i\mathbf{Z}_i^T(S) - \Sigma(S)\big)\| \geq t\big] \leq s_0 \exp\{-\frac{nt^2}{(8+4t)s_0M^2}\} \tag{10.3}$$

since $E|\mathbf{Z}_i(S)|^2 \leq s_0M^2$.

Finally,

$$\begin{aligned} P\big[\|\bar{\mathbf{Z}}\bar{\mathbf{Z}}^T(S)\|^2 \geq t\big] &= P\big[|\bar{\mathbf{Z}}|^2(S) \geq t\big] \\ &\leq s_0 P\big[|\bar{Z}_1| \geq \frac{\sqrt{t}}{s_0}\big] \leq 2s_0 e^{-\frac{tn}{2M^2s_0^2}} \ . \end{aligned} \tag{10.4}$$

By A3, (10.3) and (10.4),

$$P\big[\|\hat{\Sigma}^{-1}(S) - \Sigma^{-1}(S)\| \geq \varepsilon\big] \leq ce^{-n\delta\varepsilon^2/2} \tag{10.5}$$

for suitable c and δ. Then combining (10.2) and (10.5) the proposition follows given a consistent estimate of t. However, since we know there is a sparse representation with $\leq m_0$ factors, we can, in principle, compute $\hat{\mu}(\mathbf{Z})$ the LSE for the regression minimizing the empirical LSE on all sets of m_0 predictors. Then $\|Y - \hat{\mu}(\mathbf{Z})\|_n^2$ gives us $\hat{t}$. We can argue for consistency as we did for (10.3)-(10.5). □

Note that we could let s_0 tend to ∞ but this is of little interest. The major open question is formulating conditions under which,

(1) σ^2 (or $\|\mu(\mathbf{Z})\|^2$) is consistently estimable

(2) $\mathcal{C}(\varepsilon)$ can be identified in less than $O(p^{m_0})$ operations.

We consider a data matrix $(Z_{n\times p}, \mathbf{Y}_{n\times 1})$ and the usual linear model

$$\mathbf{Y} = Z\boldsymbol{\beta}_{p\times 1} + \mathbf{e}$$

with $E(e|\mathbf{Z}) = 0$. We assume that there is a sparse representation. We seek the set of all minimal contexts $\mathcal{C}(\varepsilon)$. We assume we are given an algorithm $\ell(Z, Y, m_0)$ which returns m_0 columns indices of Z, $S = \{i_1, \ldots, i_{m_0}\}$ such that

$$\hat{\boldsymbol{\beta}}(S) = \arg\min\left\{\|\mathbf{Y} - Z(S)\boldsymbol{\beta}\|_n^2 :\ \boldsymbol{\beta}_{m_0\times 1}\right\}.$$

$\hat{\beta}(S)$, an LSE of β, will lead to an asymptotically sparse representation of $Z\boldsymbol{\beta}$.

We sketch an algorithm for finding all minimal contexts given access to oracle $\hat{\beta}^{set}(k)$ which returns k column indices of X corresponding to k variables that have best R^2 to fit Y:

$$\hat{\beta}^{set}(k) = \arg\min_{\|\beta\|_0 \leq k} \|Y - X\beta\|_2^2 \tag{10.6}$$

In reality, as we do not have access to the oracle, let $MC(X, Y, k)$ denote a generic method which generates a single minimum context of size k and with it an estimate of the best R^2. This could be the Lasso (T1996-1), Elastic Net (ZH2005-1), SCAD (FL2001-1), MC+ (Z2010-1), FoBa (Z2011-1), or $kFoward$ proposed in Chapter 7.

Given $MC()$ we construct by an iterative method $AMC(X, Y, k, \theta_r)$ a maximal collection of minimal contexts of size k with $R^2 \geq \theta_r$, such that no context is contained in the union of all other contexts. The method scales as $O(kN(k)|MC(k)|)$, where $N(k)$ is the number of minimum contexts, and $|MC(k)|$ is the computation cost of $MC(X, Y, k)$ for short.

10.2 Algorithm

Given $MC(X, Y, k)$, consider following algorithm to recover all minimum contexts with $R^2 \geq \theta_r$:

$$FC(D)$$

if $\exists j_0$ s.t. $\sum_i D_{ij_0} = m$ **then**
  return j_0
else
  $j_0 \leftarrow$ any index s.t. $\sum_i D_{ij_0} \neq 0$
  $I_1 \leftarrow \{i : D_{ij_0} \neq 0\}$
  $I_2 \leftarrow \{1, ..., m\} - I_1$
  $I_3 \leftarrow \{j : \sum_{i \in I_1} D_{ij} \neq 0\}$
  $D_{ij} \leftarrow 0$ for $\forall i \in I_2$ and $\forall j \in I_3$
  $D' \leftarrow D_{(I_2)}$
  $L \leftarrow FC(D')$
  return $\{j_0\} \cup L$
end if

$$AMC(X, Y, k, \theta_r)$$

$J_1 \leftarrow MC(X, Y, k)$
$r \leftarrow R^2$ (R square) of $\hat{\beta}^{ols}(X, Y, J_1)$
$m \leftarrow 1$
while $r > \theta_r$ **do**
  $A \leftarrow \cup_{i=1}^{m} J_i$
  Construct $D \in \{0, 1\}^{m \times |A|}$ s.t. $D_{ij} = 1$ if and only if jth element of A is in J_i.
  for $s = 1$ to k **do**
    $L_s \leftarrow FC(D)$
    $D_{ij} \leftarrow 0$ for $\forall i$ and $\forall j \in L_s$
    $I \leftarrow A^c \cup (A \backslash L_s)$
    $J \leftarrow MC(X_I, Y, k)$
    $r \leftarrow R^2$ of $\hat{\beta}^{ols}(X, Y, J)$
    if $r > \theta_r$ **then**
      $m \leftarrow m + 1$
      $J_m \leftarrow J$
      break for
    end if
  end for
end while
return $\{J_i\}_{i=1}^{m}$

For $m = 0, 1, ...$ iteratively, suppose m minimum contexts $\{J_i\}_{i=1}^m$ have been recovered, the goal is to recover the $(m+1)$th minimum context J_{m+1}, assuming that there is at least one variable of J_{m+1} is outside the support of previous minimum contexts $A = \cup_{i=1}^m J_i$. The idea is to find a set of sets of indices $C_0 = \{L_s\}$, with the property that $L_s \cap J_i \neq \emptyset$ for all s and i, and for all possible $J_{m+1} \not\subset A$, there exists $L_{s_0} \in C_0$ s.t. $J_{m+1} \subset [p] \backslash L_{s_0}$. If $|C_0|$ is small, then apply $MC(X_{[p]\backslash L_s}, Y, k)$ for all $L_s \in C_0$ would guarantee to recover minimum context J_{m+1}, or we can conclude there is no more size k minimal context with $R^2 > \theta_r$. Since none of J_i for $i \leq m$ is included in $[p] \backslash L_s$ for all L_s, any solution other than J_{m+1} would yield a significantly worse R^2 by the definition of minimum context. Now the problem boils down to find C_0 with small size efficiently. Not surprisingly, there exists such C_0 with size exactly k. One way to find it would be for $s = 1, 2, ..., k$, recursively find L_s s.t. $L_s \cap (\cup_{l=1}^{s-1} L_l) = \emptyset$ and

$$\sum_{j \in L_s} \sum_{i=1}^{m} \mathbb{I}(j \in J_i) = m \tag{10.7}$$

For details see algorithm $FC(D)$ above.

10.3 Simulation

We simulate a multi-context toy model designed so that in the presence of large p methods based on screening single variables fail. This is similar to the simulation experiment in Chapter 9. This type of example was suggested to us by Boaz Nadler and Ya'acov Ritov. Construct standardized predictors $X \in \mathbb{R}^{n \times p}$ and response $Y \in \mathbb{R}^n$ as follows. Denote $X_i \in \mathbb{R}^n$ the ith column of X. Let $Z_1, Z_2, ...$ be i.i.d. n dimensional standard Gaussian with identity covariance matrix. We construct m minimum contexts of size k with common intersection of size s, where $k \ll p$ and $s < k$. For $w \in \mathbb{R}^m$, for each $j = 1, ..., m$, for $i = (j-1)(k-s)+1, ..., j(k-s)$, for Y_0 defined immediately later, let

$$U_j = \sum_{l=(j-1)(k-s)+1}^{j(k-s)} Z_l \tag{10.8}$$

$$V_i = w_j (Z_i - \frac{Z_i^T U_j}{\|U_j\|_2^2} U_j) + Y_0 \tag{10.9}$$

$$X_i = \frac{V_i}{\sqrt{Var(V_i)}} \tag{10.10}$$

Next let $X_{m(k-s)+1}, ..., X_p$ be entry-wise i.i.d. standard Gaussian, and let

$$Y_0 = \sum_{l=m(k-s)+1}^{(m+1)(k-s)} Z_l \tag{10.11}$$

$$Y_1 = \sum_{l=m(k-s)+1}^{m(k-s)+s} X_l \tag{10.12}$$

$$Y = SNR(Y_0 + Y_1) + \epsilon \tag{10.13}$$

where SNR is signal to noise ratio, $\epsilon \in \mathbb{R}^n$ is entry-wise i.i.d standard Gaussian independent of X. We have constructed m minimum contexts of size k with a common intersection of size $s < k$, where the jth minimum context is of the form:

$$X_{J_j} = \{X_i : i \in J_j\} \tag{10.14}$$

$$J_j = \{(j-1)(k-s)+1, ..., j(k-s)\} \cup \{m(k-s)+1, ..., m(k-s)+s\} \tag{10.15}$$

w is a parameter controlling difficulty of recovery of minimum context: if $X_i \in J_j$ is constructed with associate w_j, then $cor(X_i, Y)$ approximately scales as $\frac{k-s}{\sqrt{(k-s+w_j^2)k}}$, i.e. the larger w_j is, the weaker correlation between X_i and Y. For small w which is relatively easy to recover, every method performs more or less the same. However for bigger w, clear distinctions among methods are observed.

Next we run $AMC(X, Y, k, \theta_r)$ with $MC(X, Y, k)$. Different versions of MC include glmnet(Lasso (T1996-1) and Elastic Net (ZH2005-1)), SCAD (FL2001-1), MC+ (Z2010-1), FoBa (Z2011-1) and $kForward$ proposed in Chapter 7. All methods are used as described in Chapter 9 Section 9.1.

Specifically, an experiment is simulated for 4 minimum contexts each of size 6 with a common intersection of size 2, signal to noise ratio $SNR = 1$, sample size $n = 150$ and dimension $p = 1000, 2000, 3000$. In this case $k = 4, s = 2, m = 4$. As mentioned above, $w \in \mathbb{R}^m$ is a measurement of recovery difficulty of individual minimum context. Three choices of w are used in the simulation: $w = (5, 5, 5, 5)$, $w = (10, 10, 10, 10)$ and $w = (15, 15, 15, 15)$. For X_i constructed using w, the mean and sd of $cor(X_i, Y)$ are reported in Table 10.1. For 20 iterations, the mean and sd of number of minimum contexts(out of 4) recovered by each method are reported in Table 10.2, Figure 10.1, Figure 10.2 and Figure 10.3.

w	Mean$(cor(X_i, Y))$	SD$(cor(X_i, Y))$
5	0.316	0.077
10	0.171	0.078
15	0.122	0.082

Table 10.1: Mean and SD of $cor(X_i, Y)$ for X_i with associated w

w	p	kFoward	SCAD	MC+	FoBa	glmnet(Lasso)
5	1000	4/0	4/0	3.5/0.51	3.15/0.75	3.4/1.47
5	2000	3.9/0.31	3.95/0.22	3.3/0.73	2.75/0.79	3.6/1.23
5	3000	3.9/0.31	3.65/0.67	3.15/0.93	2.95/0.89	2.5/1.91
10	1000	4/0	3.8/0.62	2.7/1.08	3.5/0.51	3/1.59
10	2000	4/0	3.4/0.88	1.65/1.35	2.8/1.11	2/1.65
10	3000	4/0	2.8/1.06	0.95/1.19	1.6/1.43	0.15/0.67
15	1000	4/0	1.65/1.28	0/0	1.85/0.97	0/0
15	2000	4/0	0.8/0.73	0/0	0.5/0.76	0/0
15	3000	4/0	0.2/0.62	0/0	0/0	0/0

Table 10.2: Mean/SD of number of recovered minimum contexts (out of 4) for 20 iterations

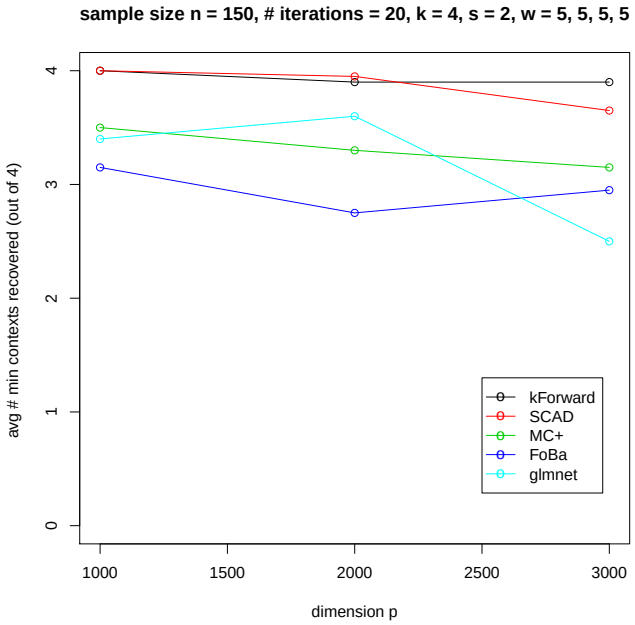

Figure 10.1: Average number of minimum contexts recovered for $w = (5, 5, 5, 5)$.

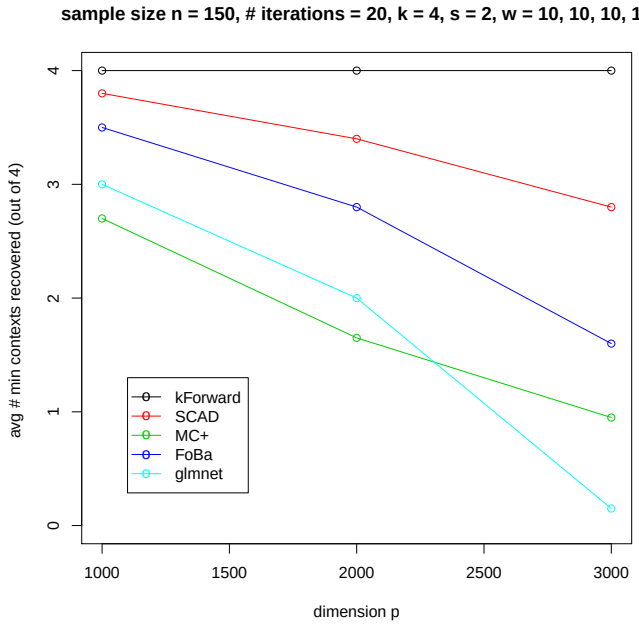

Figure 10.2: Average number of minimum contexts recovered for $w = (10, 10, 10, 10)$.

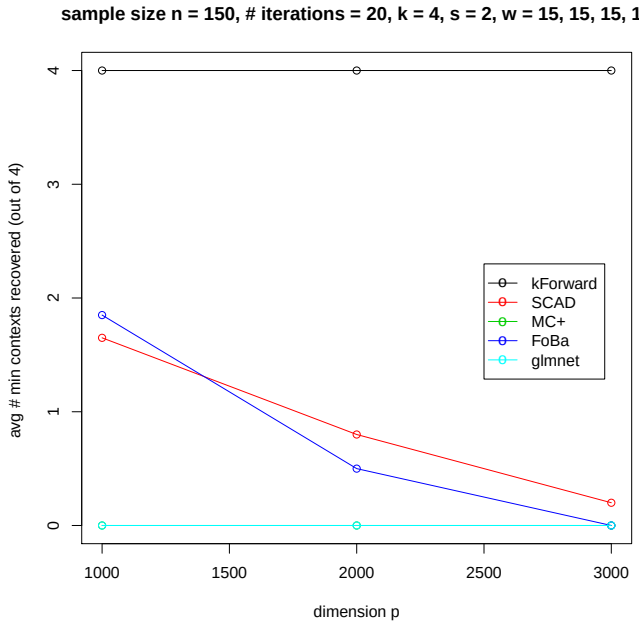

Figure 10.3: Average number of minimum contexts recovered for $w = (15, 15, 15, 15)$.

Bibliography

BL2008a

Peter J. Bickel and Elizaveta Levina, Regularized estimation of large covariance matrices, Annals of Statistics, 2008.

BL2008b

Peter J. Bickel and Elizaveta Levina, Covariance regularization by thresholding, Annals of Statistics, 2008.

CZZ2010

T. Tony Cai, Cun-Hui Zhang and Harrison H. Zhou, Optimal rates of convergence for covariance matrix estimation, Annals of Statistics, 2010.

CL2011

T. Cai and W. Liu, Adaptive thresholding for sparse covariance matrix estimation, Journal of American Statistical Association, 2011.

CZ2012

T. Tony Cai and Harrison H. Zhou, Optimal rates of convergence for sparse covariance matrix estimation, Annals of Statistics, 2012.

CY2012

T. Tony Cai and Ming Yuan, Adaptive covariance matrix estimation through block thresholding, Annals of Statistics, 2012.

EK2007

Noureddine El Karoui, Operator norm consistent estimation of large dimensional sparse covariance matrices, Annals of Statistics, 2007.

FB2007

R. Furrer and T. Bengtsson, Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants. J. Multivariate Anal, 2007.

AW2009

Arash A. Amini and M. J. Wainwright, High-dimensional analysis of semidefinite programming relaxations for sparse principal component analysis, Annals of Statistics, 2009.

J2001

I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Annals of Statistics, 2001.

DA2007

A. D'Aspremont, L. El Ghaoui, M.I. Jordan and G.R.G. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, SIAM Rev, 2007.

ZHT2006

H. Zou, T. Hastie and R. Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics, 2006.

J2003

I.T. Jolliffe, N.T. Rendafilov and M. Uddin, A modified principal component technique based on the LASSO, Journal of Computational and Graphical Statistics, 2003.

BL2010

Peter J. Bickel and Marko Lindner, Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics, arXiv:1002.4545v2, 2010.

FHT2008

Jerome Friedman, Trevor Hastie and Robert Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 2008.

R2008

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu, Sparse permutation invariant covariance estimation, Electronic Journal of Statistics, 2008.

T1996

R. Tibshirani, Regression shrinkage and selection via the lasso. Journal of Royal Statistics Society B, 1996.

YZ2011

Xiaotong Yuan and Tong Zhang, Truncated Power Method for Sparse Eigenvalue Problems, arXiv:1112.2679, 2011.

Y1997

Bin Yu, Assouad, Fano, and Le Cam, Festschrift for Lucien Le Cam. D. Pollard, E. Torgersen, and G. Yang, pp. 423-435, Springer-Verla, 1997.

D1992

D.L. Donoho, I.M. Johnstone, J.C. Hoch and A.S. Stern, Maximum entropy and the nearly black object, Journal of the Royal Statistical Society, Series B, 1992.

CDS1998

S.S. Chen, D.L. Donoho and M.A. Saunders, Atomic Decomposition by Basis Pursuit, SIAM Journal on scientific computing, 1998.

MB2006

Nicolai Meinshausen and Peter B*ü*hlmann, High-dimensional graphs and variable selection with the Lasso, Annals of Statistics, 2006.

ZY2006

Peng Zhao and Bin Yu, On Model Selection Consistency of Lasso, The Journal of Machine Learning Research, 2006.

W2006

M. Wainwright, Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical Report 708, Dept. Statistics, Univ. California, Berkeley, 2006.

CT2007

E. Cand*é*s and T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n (with discussion), Annals of Statistics, 2007.

BRT2009

Peter J. Bickel, Ya'acov Ritov and Alexandre B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Annals of Statistics, 2009.

ZH2005

H. Zou and T. Hastie, Regularization and Variable Selection via the Elastic Net, Journal of the Royal Statistical Society, Series B, 2005.

Z2006

H. Zou, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, 2006.

FL2001

J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and it oracle properties, Journal of American Statistical Association, 2001.

FL2002

J. Fan and R. Li, Variable Selection for Cox's Proportional Hazards Model and Frailty Model, Annals of Statistics, 2002.

Z2010

Cun-Hui Zhang, Nearly unbiased variable selection under minimax concave penalty, Annals of Statistics, 2010.

Z2011

T. Zhang, Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations, IEEE Transactions on Information Theory, 2011.

B2011

Richard A. Berk, Lawrence D. Brown, Andreas Buja, Kai Zhang and Linda Zhao, Valid Post-Selection Inference, Annals of Statistics, 2011.
FK2000

Wenjiang Fu and Keith Knight, Asymptotics for lasso-type estimators, Annals of Statistics, 2000.

O2010

Roberto Imbuzeiro Oliveira, Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges, arXiv:0911.0600, 2010.
J2012

Jiashun Jin, Cun-Hui Zhang and Qi Zhang, Optimality of Graphlet Screening in High Dimensional Variable Selection, arXiv: 1204.6452, 2012.
K2012

Tracy Ke, Jiashun Jin and Jianqing Fan, Covariance Assited Screening and Estimation, arXiv: 1205.4645, 2012.
S2011

Liangcai Shu, Aiyou Chen, Ming Xiong and Weiyi Meng, Efficient Spectral Neighborhood Blocking for Entity Resolution, ICDE 2011.
P2002

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser, Identity uncertainty and citation matching, Advances in Neural Information Processing (NIPS), 2002.
F2009

W. Fan, X. Jia, J. Li, and S. Ma, Reasoning about record matching rules, The 35th International Conference on Very Large Data Bases (VLDB), 2009.
N1959

H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, Automatic linkage of vital records, Science, 1959.
F1969

I. Fellegi and A. Sunter, A theory for record linkage, Journal of the American Statistical Society, 1969.
B2004

I. Bhattacharya and L. Getoor, Iterative record linkage for cleaning and integration, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004.
B2003

M. Bilenko and R. J. Mooney, Adaptive duplicate detection using learnable string similarity measures, SIGKDD, 2003.
BG2004

I. Bhattacharya and L. Getoor, Deduplication and group detection using links, ACM SIGKDD Workshop on Link Analysis and Group Detection, 2004.