Group Sparsity in Regression and PCA

by

Yanzhen Deng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Tailen Hsing, Chair
Professor Alfred Hero
Associate Professor Ambuj Tewari
Professor Ji Zhu

Yanzhen Deng

dengyz@umich.edu

ORCID iD: 0000-0003-0213-7994

ACKNOWLEDGEMENTS

First of all, no word can express my gratitude to Professor Tailen Hsing, for his advices on both my research and my life. Without his patient guidance and continuous support, I will be long lost in frustration and never finish my Ph.D. study.

Second, I want to thank Professor Ji Zhu, Associate Professor Ambuj Tewari and Professor Alfred Hero for being on my committee.

Also, I want thank Professor Shuheng Zhou for her instruction in the early stage. Also special thank to Professor Stilian Stoev for his suggestion on some proofs in the thesis.

Finally, I want to thank my family, especially my wife, for cheering me up when I lose faith in myself.

ABSTRACT

In the field of high-dimensional statistics, it is commonly assumed that only a small subset of the variables are relevant and sparse estimators are pursued to exploit this assumption. Sparse estimation methodologies are often straightforward to construct, and indeed there is a full spectrum of sparse algorithms covering almost all statistical learning problems. In contrast, theoretical developments are more limited and often focus on asymptotic theories. In applications, non-asymptotic results may be more relevant.

The goal of this work is to show how non-asymptotic statistical theory can be developed for sparse estimation problems that assume group sparsity. We discuss three different problems: principal component analysis (PCA), sliced inverse regression (SIR) and multivariate regression. For PCA, we study a two-stage thresholding algorithm and provide theories that go beyond the common spiked-covariance model. SIR is then related to PCA in some special settings, and it is shown that the theory of sparse PCA can be modified to work for SIR. Regression represents another important research direction in high-dimensional analysis. We study a linear regression model in which both the response and predictors are grouped, as an extension of group Lasso.

Despite the distinctions in these problems, the proofs of consistency and support recovery share some common elements: concentration inequalities and union probability bounds, which are also the foundation of most existing sparse estimation theories. The proofs are presented in modules in order to clearly reveal how most sparse estimators can be theoretically justified. Moreover, we identify those modules that are possibly not optimized to show the limitation of the existing proof techniques and how they could be extended.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . **ii**

ABSTRACT . . . . **iii**

LIST OF FIGURES . . . . **vii**

LIST OF TABLES . . . . **viii**

CHAPTER

I. Introduction 1

1.1 Overview 1
1.2 Contribution 3
1.3 Notation 3
1.3.1 Slicing 3
1.3.2 Matrix and vector norms 4
1.3.3 Miscellaneous 5

II. Principal Component Analysis with Group Sparsity 6

2.1 Introduction 6
2.1.1 Overview of Sparse PCA(SPCA) 6
2.1.2 Group Sparsity in PCA 8
2.1.3 Outline and setup 10
2.2 SPCA based on penalization 11
2.2.1 SPCA via regression formulation (SPCA-reg) 11
2.2.2 SPCA with soft-thresholded power method (SPCA-Power) 13
2.2.3 SPCA via semidefinite programming (SPCA-SDP) 15
2.3 SPCA based on thresholding 18
2.3.1 SPCA with diagonal thresholding (SPCA-DT) 18
2.3.2 SPCA with augmented thresholding (SPCA-AT) 19
2.4 Statistical property of SPCA-DT and SPCA-AT 21
2.4.1 Distance between linear subspaces 21
2.4.2 Single-spike and i.i.d. noise, $D = 1$, $\Sigma_e = \sigma_e^2 \mathbf{I}_p$ 23
2.4.3 Single-spike and non-i.i.d. noises, $D = 1$, $\Sigma_e \neq \sigma_e^2 \mathbf{I}_p$ 35
2.4.4 Multiple spikes $D > 1$ 45
2.5 Simulation 54
2.6 Auxiliary lemmas 58
2.6.1 Exponential tail bounds 58
2.6.2 Eigenvalue and eigenvector perturbation 61

III. Sliced Inverse Regression with Group Sparsity 64

3.1 Introduction . . . 64
3.2 Sparse SIR methods . . . 67
3.2.1 SSIR using SPCA-SDP . . . 68
3.2.2 SSIR via Lasso . . . 70
3.2.3 Thresholding-based SSIR . . . 71
3.2.4 Idea of refinement . . . 72
3.3 Statistical property . . . 73
3.3.1 Special case: Gaussian linear model . . . 74
3.3.2 General model, $D = 1$, $\Sigma_X = \mathbf{I}_p$ . . . 80
3.3.3 Discussions . . . 84
3.4 Simulation on SSIR-DT/AT . . . 85
3.4.1 Asymptotic rate of sample size . . . 85
3.4.2 Choice of slice number H . . . 93
3.5 Comparison of SSIR methods . . . 99

IV. Regression with Block Sparsity . . . 101

4.1 Introduction . . . 101
4.2 Statistical consistency . . . 104
4.2.1 Notation . . . 104
4.2.2 Exact sparsity recovery . . . 104
4.2.3 Lower bound of λ_n . . . 119
4.2.4 Compare to Lasso and group Lasso . . . 121
4.2.5 Disregard block structure . . . 122
4.3 Numerical study . . . 124
4.4 Appendix . . . 127
4.4.1 Auxiliary lemmas . . . 127
4.5 Numerical study on $\left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\right\|_{\infty,F}$ . . . 133

V. Future Research Directions . . . 136

BIBLOGRAPHY . . . **138**

LIST OF FIGURES

Figure

2.1 Probability of estimating the correct sparse set 20

3.1 Model I, Σ_1 87

3.2 Model I, Σ_2 88

3.3 Model II, Σ_1 88

3.4 Model II, Σ_2 88

3.5 Model III, Σ_1 89

3.6 Model III, Σ_2 89

3.7 Model IV, Σ_1 89

3.8 Model IV, Σ_2 90

3.9 Model V, Σ_1 90

3.10 Model V, Σ_2 90

3.11 Rows corresponding to Model I to V, respectively; $\Sigma_X = \Sigma_1$ 91

3.12 Rows corresponding to Model I to V, respectively; $\Sigma_X = \Sigma_2$ 92

3.13 Model I, top two rows: Σ_1, bottom two rows: Σ_2. 94

3.14 Model II, top two rows: Σ_1, bottom two rows: Σ_2. 95

3.15 Model III, top two rows: Σ_1, bottom two rows: Σ_2. 96

3.16 Model IV, top two rows: Σ_1, bottom two rows: Σ_2. 97

3.17 Model V, top two rows: Σ_1, bottom two rows: Σ_2. 98

4.1 Empirical probability of exact recovery against raw sample size vs. scaled sample size, logarithmic sparsity 126

4.2 Empirical probability of exact recovery against raw sample size vs. scaled sample size, linear sparsity 126

4.3 The distribution of $\left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\right\|_{\infty,F}$ when s increases. 134

LIST OF TABLES

Table

2.1 Compare the performance of various SPCA algorithm under a wide range of settings. 56

2.2 Trade off between sample size and signal. $n\rho^2 = 4000$. 57

2.3 Performance when p increases 57

3.1 Comparison between various SSIR methods. 100

CHAPTER I

Introduction

1.1 Overview

We study various statistical methods for group sparsity, i.e. sparsity defined on groups of coefficients and not on individuals. A typical example is multi-variate linear regression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}, \boldsymbol{\epsilon} \in \mathbb{R}^{n\times D}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^{p\times D}$. If we know that only a small subset of X_d's are relevant in the model, then $\boldsymbol{\beta}$ has many zero rows, or row-wise sparse. An interpretable estimator should ideally also be row-wise sparse, so that it achieves variable selection simultaneously. One can simply regress each dimension of the response variables separately, and make estimator of the columns $\boldsymbol{\beta}_{\cdot d}$ sparse. Some well-studied estimators like the Lasso estimator or the Dantzig selector can be used. The implementation of these estimators can be easily carried out by ℓ_1 linear programming, and well-established packages are available. However, without any ad hoc treatment, "common support" is not guaranteed, in which case the variables selected are the union of the supports of all of the D eigenvectors. Also, one needs to tune the D sub-problems individually, which adds complication and may cause over-tuning.

Alternatively, the goal can be recovering the row-wise support directly. In this regard, instead of running D Lasso regression, one can apply a group-Lasso penalization, $\hat{\boldsymbol{\beta}} = \arg\min \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_{\mathrm{F}}^2 + \lambda \sum_{j=1}^{p} \left\| \boldsymbol{\beta}_{j\cdot} \right\|_2$. By tuning λ, the elements in the same row of $\hat{\boldsymbol{\beta}}$ will be zero or non-zero simultaneously. In regard to the group-wise sparse assumption (that only a few variables X_i's are relevant), this estimator seems to be more suitable. However, generally speaking, the group sparsity assumption presents more challenges both computationally and theoretically.

Many statistical problems have a group-wise sparse version. In linear regression, two closely related topics are group Lasso (Yuan and Lin 2006, Nardi and Rinaldo 2008, Bach 2008, Huang and Zhang 2010) and sparse multivariate-response regression (Karim Lounici and Tsybakov 2011, Meier et al. 2009, Obozinski et al. 2011, Negahban et al. 2012, Liu and Zhang 2008). The fundamental idea behind these works is to use some group-wise norm of the coefficients as the penalty so that coefficients in the same group shrink to 0 simultaneously. For instance, in the principal component analysis (PCA) literature, Johnstone and Lu (2009), Birnbaum et al. (2013), Ma (2013) use thresholding to screen variables so that the resulting PC loadings share the same sparsity. A lot of works on the other hand apply regularization-based approaches to achieve sparsity, e.g. Jolliffe et al. (2003), Zou and Xue (2006), Journée et al. (2010), Witten et al. (2009), Shen and Huang (2008), d'Aspremont et al. (2005). It is common that PC's are estimated sequentially so that common support is not guaranteed, but usually the algorithm can be modified to achieve common support. Techniques that "sparsify" PCA usually can also be applied to other eigenvalue problems like generalized eigenvalue problem (GEP) and canonical correlation analysis (CCA). In sliced inverse regression (SIR), recent development also considered common support, e.g. Zhu et al. (2011), Lin et al. (2018), Tan et al. (2017).

1.2 Contribution

In this thesis, we investigate three topics: sparse PCA, sparse SIR and block-wise penalized regression. Our main contribution is to demonstrate how to construct non-asymptotic statistical theories for sparse estimators that assume group sparsity. The proofs of all these results share two common elements: concentration inequalities and union probability bounds, which are also the basis of most approaches in the sparse estimation literature. When converted to asymptotic theories, our results are either comparable or better than those in the literature.

Another contribution is the clarification of the proof structure. We present the proofs in modules so that, with suitable modifications, they would work for different model assumptions than those assumed in this work. Also, we can easily identify the modules that are possibly not optimized to show the limitation of our proof techniques.

Finally, we make some interesting observations from our numerical studies. For sparse PCA, our simulations show that thresholding-based methods actually perform as well as regularization-based methods, even though the ideas are naive. For sparse SIR, we observe that the slice size is an important factor that affects the performance of the approach.

1.3 Notation

1.3.1 Slicing

Since we deal with groups and blocks of elements a lot, we need to first set up a notation system of slicing.

For vector $v \in \mathbb{R}^n$, we use v_i to denote its i-th element; for a subset of indices $I = \{i_1, ..., i_s\} \subseteq \{1, 2, ..., n\}$, v_I is the sub-vector $(v_{i_1}, ..., v_{i_s})$. When there is a clear grouping of elements $\{1, 2, ..., n\} = \uplus_{j=1}^{p} G_j$, we may also use $v_{[j]}$ to denote v_{G_j}.

For matrix $M \in \mathbb{R}^{n \times m}$, we use two subscripts to indicate row index and column index, respectively. For example, M_{ij} is the element at the i-th row and the j-th column, where one or both of the subscripts can refer to subsets; for example if $I = \{1, 2\}$, then $M_{1I} = \begin{pmatrix} M_{11} & M_{12} \end{pmatrix}$, $M_{I1} = \begin{pmatrix} M_{11} \\ M_{21} \end{pmatrix}$, and $M_{II} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$. If we want to select all elements in one row/column or a subset of rows/columns, then we use "$\cdot$" to stand for full set, so for example, $M_{1\cdot} = \begin{pmatrix} M_1 & \cdots & M_{1m} \end{pmatrix}$, $M_{\cdot I} = \begin{pmatrix} M_{11} & M_{21} \\ \vdots & \vdots \\ M_{n1} & M_{n2} \end{pmatrix}$. When the matrix is partitioned in rows $\{1, 2, ..., n\} = \uplus_{j=1}^{p} G_j$, then we denote $M_{[j]} = M_{G_j \cdot}$; if it is partitioned in columns $\{1, 2, ..., m\} = \uplus_{j=1}^{p} H_k$, then $M_{[k]} = M_{\cdot H_k}$. If the matrix is partitioned two ways, so that $\{1, 2, ..., n\} = \uplus_{j=1}^{p} G_j$ and $\{1, 2, ..., m\} = \uplus_{k=1}^{l} H_k$, then we use $M_{[jk]}$, $M_{[j\cdot]}$ and $M_{[\cdot k]}$ to indicate a single block, row of blocks and column of blocks. We might also include several rows or columns of blocks. For example, if $I = \{1, 2\}$, then $M_{[II]} = \begin{pmatrix} M_{[11]} & M_{[12]} \\ M_{[21]} & M_{[22]} \end{pmatrix}$.

1.3.2 Matrix and vector norms

For a vector $v = (v_1, ..., v_p)$, we denote $\|v\|_0 = \#\{i : v_i \neq 0\}$, and $\|v\|_q = (\sum_{i=1}^{n} |v_i|^q)^{1/q}$ for $q > 0$. Some common examples are $\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$ and $\|v\|_1 = \sum_{i=1}^{n} |v_i|$. If the elements of v are grouped into p groups $\{1, 2, ..., n\} = \uplus_{j=1}^{p} G_j$, then we define $\|v\|_{p,q} = \left\| \left(\|v_{G_1}\|_q, \cdots, \|v_{G_p}\|_q \right) \right\|_p$. A common example is the group Lasso penalty, $\|v\|_{1,2} = \sum_{j=1}^{p} \|v_{G_j}\|_2$.

For a matrix $M \in \mathbb{R}^{n \times m}$, we denote by $\sigma_{\min}(M)$, $\sigma_{\max}(M)$ the minimal and maximal singular values of M. Denote the operator norm of M by $\|M\|_{\mathrm{op}} = \max \frac{\|Mv\|_{\mathrm{op}}}{\|v\|_{\mathrm{op}}}$, which is just $\sigma_{\max}(M)$, but is used when M is a positive semi-definite matrix. Denote Frobenius norm $\|F\|_{\mathrm{F}}^2 = \sum_{i,j} M_{ij}^2$, and nuclear norm $\|M\|_* = \sum_{d=1}^{m\wedge n} \sigma_d(M)$, where $\sigma_d(M)$ is the d-th singular value of M. We use $\|M\|_{\infty\to\infty}$ to denote operator norm w.r.t. ℓ_∞ vector norm.

We use two subscripts to denote for group-wise or block-wise norm. The second subscript indicates the norm applied to sub-groups/blocks; the first subscript indicates the norm applied to the norms of sub-groups/blocks. For example, we use $\|M\|_{1,2} = \sum_{j=1}^{m} \|M_{j\cdot}\|_2$ stands for the sum of row norms; if rows of M are divided in to p groups $\{1, 2, ..., n\} = \uplus_{j=1}^{p} G_j$, then $\|M\|_{1,F} = \sum_{j=1}^{p} \left\|M_{G_j\cdot}\right\|_{\mathrm{F}}$ stands for the sum of block norms.

1.3.3 Miscellaneous

Denote zero vector and one vector by $\mathbf{0}_n = (0, ..., 0)$ and $\mathbf{1}_n = (1, ..., 1)$, respectively; let $\mathbf{0}_{n\times m}, \mathbf{1}_{n\times m}$ be n by m matrices with all zero and one elements, respectively. The diagonal matrix with diagonal elements $v = (v_1, v_2, ...)$ is denoted by $M = \mathbf{diag}(v)$. The identity matrix of size p is denoted by $\mathbf{I}_n = \mathbf{diag}(\mathbf{1}_n)$. Denote the Kronecker product of two matrices by $M_1 \otimes M_2$.

For a vector $v = (v_1, ..., v_n)$ and a $c > 0$, denote soft-thresholding operator by

$$\mathcal{T}(v, c) = (\mathbf{sign}(v_1 - c)(v_1 - c)_+, ..., \mathbf{sign}(v_n - c)(v_n - c)_+)$$

and let $\mathcal{T}(M, c)$ be the element-wise soft-thresholding on matrix M.

We also set asymptotic notation of sequences. Let $\{a_n\}$ and $\{b_n\}$ be two sequences of numbers. We say $a_n \succeq b_n$ if there exists $C > 0$ so that $|a_n| > C|b_n|$ when n is large enough; if for any $C > 0$, $|a_n| > C|b_n|$ when n is large enough, then we say $a_n \succ b_n$; if there exists $0 < c < C$ so that $c|b_n| < |a_n| < C|b_n|$ when n is large enough, then $a_n \asymp b_n$; if $b_n \succeq (\succ) a_n$, then $a_n \preceq (\prec) b_n$.

CHAPTER II

Principal Component Analysis with Group Sparsity

2.1 Introduction

2.1.1 Overview of Sparse PCA(SPCA)

PCA is widely applied for dimension reduction, which is especially useful for high-dimensional data. The ordinary PCA is equivalent to eigenvalue decomposition and singular value decomposition. The leading PC loadings of a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ are just the leading right singular vectors, or leading eigenvectors of $\mathbf{X}^T\mathbf{X}/n$. In the rest of the work, we will use the terms: loadings, eigenvectors, leading directions interchangeably.

There are two main reasons why a sparse version of PCA is desirable. The first reason is interpretability. When p is large, the loadings are hard to interpret because they are simply vectors full of small non-zero numbers. It is preferable to have sparse loadings because it would be easier to determine which variables are important for explaining the variability.

Another reason, which is more important in our opinion, is consistency. It is well-known that sample eigenvectors are generally poor estimators of the population eigenvectors when dimension is truly high. To be more specific, p/n has to converge to 0 in order for the sample eigenvectors to be consistent. The inconsistency phenomenon occurs not only when $\mathbf{X}$ are samples from multivariate Gaussian distribution, but also occurs for other model setups. Some regularization is thus needed in order for PCA to be actually useful. Even if sparsity is not intended, a sparse estimator may still be better than ordinary PCA.

In order to get sparse eigenvectors, there are two approaches: penalization and thresholding. The penalization approach is to formulate PCA as some optimization problem and then add a sparsity-inducing penalty (or constraint, which usually has an equivalent penalization form). In the literature, a common approach is ℓ_1 penalty/constraint as a relaxation of ℓ_0 constraint. For example, in an early work, Jolliffe et al. (2003) propose ScotLass. To get the top eigenvector, it solves the following optimization problem

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\textbf{maximize:}} \quad & v^T\mathbf{X}^T\mathbf{X}v \\ \textbf{subject to:} \quad & \|v\|_2 \le 1, \|v\|_1 \le \tau \ . \end{aligned}$$

This is a natural way to "sparsify" eigenvectors. Without the ℓ_1 sparsity constraint, the solution is exactly the top eigenvector. ScotLass is inspired by Lasso, but the computation complexity is high because it is not convex. A lot of follow-up works aim to provide efficient algorithms to approximately solve it. Basically, instead of solving a non-convex optimization like in ScotLass, one can penalize some equivalent forms that are still easy to compute after adding the ℓ_1 penalty.

The thresholding approach is to first screen the variables, obtain the eigenvectors based on the included subset, and make the loadings of the excluded variables zero. By thresholding, solving eigenvectors of a large matrix is totally avoided, so the computation is usually much lighter then penalization-based approaches.

The easiest thresholding-based method is diagonal thresholding (DTSPCA) proposed by Johnstone and Lu (2009). The idea is to exclude variables whose marginal sample variances are small. This method is crude because the covariance information is not used. Some follow-up works proposed extra thresholding steps to improve the performance and to achieve near-optimal error rate. Examples include augmented SPCA (ASPCA) proposed by Birnbaum et al. (2013) and iterative thresholding (ITSPCA) proposed by Ma (2013).

Various theoretical results have been derived. Zou et al. (2018) provided a review on SPCA and a comprehensive list of references on the theoretical aspects of SPCA. It is not always possible to exactly compare different theoretical results because they are based on various model assumptions. Some typical technical differences are: exact sparsity $\|v\|_0 \leq s$ versus approximate sparsity $\|v\|_q \leq s$ for some $0 < q < 1$; fixed rank D versus D diverging with sample size n; Gaussian distribution versus non-Gaussian (typically sub-Gaussian) distribution; whether n or signal-to-noise ratio ρ goes to infinity, and so on. There are three types of theories that can be of interest: minimax lower bounds, upper bounds of estimation error and so-called computational lower bounds. Essentially, the minimax error rate of the estimated eigenvectors is $\frac{s\log(p/s)}{n\rho}$, while the upper bound is of rate $\frac{s\log p}{n\rho}$. A computational lower bound is the minimal sample size for consistency. Even if we prove the upper bound of error is of rate $\frac{s\log p}{n\rho}$, having $n \succ \frac{s\log p}{\rho}$ may not be enough for consistency, because sometimes the error bound needs an extra sample size assumption to hold. Typically, it is assumed $\frac{s^2\log p}{n\rho} = O(1)$, and several works have proved that $n \succ \frac{s^{2-\delta}\log p}{\rho}$ is necessary.

2.1.2 Group Sparsity in PCA

Often, the top PC is not enough to represent the data. To get subsequent sparse eigenvectors is less straight forward than getting the top sparse eigenvectors, especially for penalization-based methods. There are two potential choices that one can make when constructing an algorithm.

The first choice, which is relevant to our main topic, is whether to control the sparsity of multiple eigenvectors separately or jointly. If $\hat{\mathbf{V}} = (\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_D)$ are the estimated leading eigenvectors, then one may want to control common support $\{k : \hat{\mathbf{v}}_{dk} \neq 0 \ \exists d \leq D\}$. We call this group-wise sparse PCA (GSPCA). Note that thresholding-based approaches naturally achieve common support, so GSPCA is the same as SPCA for thresholding; besides, if only the top PC is sought, then GSPCA reduces to SPCA. Thus, group sparsity is non-trivial when using a penalization-based approach to go beyond top PC. In the literature of SPCA, group sparsity is rarely considered, so most penalized SPCA algorithms need to be modified to serve as GSPCA algorithms.

The second choice, is whether the target of the algorithm is the actual eigenvectors $(\mathbf{v}_1, ..., \mathbf{v}_D)$ or the leading principal space (or eigen-space) $\mathcal{V} = \mathbf{span}\langle \mathbf{v}_1, ..., \mathbf{v}_D \rangle$. If we only care how much information is retained by the top D PC's, then the eigen-space is good enough, because $\mathbf{span}\langle \mathbf{v}_1, ..., \mathbf{v}_D \rangle = \mathbf{span}\langle \mathbf{v}'_1, ..., \mathbf{v}'_D \rangle$ means $(\mathbf{X}\mathbf{v}_1, ...\mathbf{X}\mathbf{v}_D)$ and $(\mathbf{X}\mathbf{v}'_1, ..., \mathbf{X}\mathbf{v}'_D)$ contain equivalent information. This is often enough if the purpose is dimension reduction.

These two choices are connected with each other. If the individual eigenvectors are of interest, and the goal is to estimate the actual eigenvectors, then it is quite inevitable to use some sequential style algorithms. Typically, the idea is to first estimate the top eigenvector $\hat{\mathbf{v}}_1$; then when estimating the next eigenvector $\hat{\mathbf{v}}_2$, one can apply an extra restriction $\hat{\mathbf{v}}_1^T \hat{\mathbf{v}}_2 = 0$; another option is to obtain the top eigenvector of $\mathbf{X}(\mathbf{I}_p - \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T)$ as $\hat{\mathbf{v}}_2 = 0$. Note that it is hard to control the estimated eigenvectors to have common support with this type of algorithm,

If the individual eigenvectors are not of interest, and the goal is just to estimate the leading eigen-space, then one can estimate all eigenvectors together by solving one optimization. In that way it is easier to get common support. For example, with the same rationale as group Lasso, one may add $\|V\|_{1,2} = \sum_{j=1}^{p} \|V_{j\cdot}\|_2$ as penalty to $\min_{V^TV=\mathbf{I}_p} -\mathbf{tr}\left(V^T\mathbf{X}^T\mathbf{X}V\right)$. It is easy to see that without ℓ_1 penalty, the algorithm does not even have a unique solution. Indeed, the objective is summation of the variance of the leading PC's; any orthonormal transformation VO where $O^TO = \mathbf{I}_D$ will not change the objective at all. Thus, with penalty added, one should not expect that estimator is close to the true eigenvectors, but only close to some linear transformation of the eigenvectors. Thus, such algorithm is only valid if the target is the eigen-space. For example, if we have $p = 100$ variables, but we want to reduce to $D = 3$ using only $s = 10$ of them, then our goal falls into this category.

Note that if a thresholding-based approach is used, then naturally the estimated vectors that share the same support, so the above differentiation is irrelevant.

2.1.3 Outline and setup

Here is the layout of this chapter. In section 2.2 we review some typical penalization-based SPCA algorithms. We will modify these algorithms so the estimated PC loadings share the same support, thus achieving the model selection purpose more efficiently. In section 2.3 we will specify two thresholded SPCA algorithms, and in section 2.4 we provide some non-asymptotic error bounds of the estimated leading space using thresholded SPCA. These results are the foundation for the more complicated SIR problem. Finally, in section 2.5 we run some numerical studies to compare different SPCA algorithms.

Throughout this chapter, we assume that the data matrix $\mathbf{X}$ is generated from a "fixed-design" spiked model, that is

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T + \boldsymbol{\epsilon}\ ,$$

where $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_D) \in \mathbb{R}^{n\times D}$ and $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_D) \in \mathbb{R}^{p\times D}$ satisfies $\frac{\mathbf{U}^T\mathbf{U}}{n} =$

$\mathbf{V}^T\mathbf{V} = \mathbf{I}_D$. $\mathbf{S} = \mathbf{diag}(\sigma_1, ..., \sigma_D)$ are the population singular values. For convenience, we denote $\mathbf{\Lambda} = \mathbf{S}^2$, $\boldsymbol{\lambda}_d = \sigma_d^2$. Conceptually $\mathbf{USV}^T$ is the low rank signal matrix, contaminated by noise $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times p}$ which we assume to have i.i.d. rows, $\epsilon_{i\cdot} \sim \mathbf{Norm}(0, \Sigma_E)$. The goal is to estimate the population leading principal space, denoted by $\mathcal{V} = \mathbf{colspan}\langle\mathbf{V}\rangle$. For simplicity, throughout this work we assume that $\mathbf{1}_n^T\mathbf{U} = 0$ so that we never need to worry about centering.

We make this model assumption so that it can be used in analysis of the sparse sliced inverse regression (SSIR) in the next Chapter. Traditionally, $\mathbf{U}$ is assumted to be standard Gaussian ensemble that is independent of $\boldsymbol{\epsilon}$. Here we instead assume $\mathbf{U}$ to be a fixed matrix such that $\mathbf{U}^T\mathbf{U}/n = \mathbf{I}_D$, which is actually more general, and theory can be easily adapted to any random $\mathbf{U}$ using some conditional arguments. We first assume $\boldsymbol{\epsilon}$ to be i.i.d. Gaussian ensemble, that is $\Sigma_E = \sigma_e^2\mathbf{I}_p$; then we relax this assumption to the non i.i.d. noise.

2.2 SPCA based on penalization

2.2.1 SPCA via regression formulation (SPCA-reg)

The first efficient SPCA algorithm was proposed by Zou and Xue (2006). It was directly called SPCA in the original work, but to avoid ambiguity, we name it SPCA-reg. Consider the "minimal residual variance" formulation of PCA, that is

$$\begin{aligned} \textbf{minimize:} \quad & \frac{1}{n}\left\|\mathbf{X}(\mathbf{I}_p - VV^T)\right\|_{\mathrm{F}}^2 \\ \textbf{subject to:} \quad & V^TV = \mathbf{I}_D\ . \end{aligned}$$

Note that the optimal solution is not unique. The sample PC loading vectors $\hat{V}_{load}$ is one of the solutions; any right orthonormal transformations of $\hat{V}_{load}$ are also solutions.

Obviously, we can introduce a duplicate argument U, and the above problem is equivalent to the following

$$\begin{aligned} \textbf{minimize:} \quad & \frac{1}{n} \left\|\mathbf{X}(\mathbf{I}_p - VU^T)\right\|_{\mathrm{F}}^2 \\ \textbf{subject to:} \quad & U^T U = \mathbf{I}_D, U = V \ . \end{aligned}$$

Note that by removing the equality constraint $U = V$, the estimated leading space will not change. More precisely, if $(\hat{U}, \hat{V})$ is one solution of

$$\begin{aligned} \textbf{minimize:} \quad & \frac{1}{n} \left\|\mathbf{X}(\mathbf{I}_p - VU^T)\right\|_{\mathrm{F}}^2 \\ \textbf{subject to:} \quad & U^T U = \mathbf{I}_D \ . \end{aligned}$$

then there exists an orthonormal matrix $O \in \mathbb{R}^{D\times D}$, and some constant $c_1, ..., c_D$, such that $\hat{U}O = \hat{V}_{load}$, and $\hat{V} = \hat{V}_{load}\mathbf{diag}(c_1, ..., c_D)O$. Therefore $\mathbf{colspan}\langle\hat{V}\rangle = \mathbf{colspan}\langle\hat{V}_{load}\rangle$ is the sample leading space. It can also be proved that the estimated leading space will not change by adding a ridge penalty to the objective function.

Remark II.1. *In the original paper (Zou and Xue 2006, Theorem 3), the author stated that columns of non-sparse-penalized* $\hat{V}$ *are parallel to the columns of* $\hat{V}_{load}$*, i.e.,* $\hat{V} = \hat{V}_{load}\mathbf{diag}(c_1, ..., c_D)$ *for some* $c_1, ..., c_D$*. Thus after getting the sparse-penalized* $\hat{V}$*, they normalize by columns to get estimator of the actual PC loading vectors. This is incorrect, as the optimization is not targeting the eigenvectors, but the linear space spanned by them.*

Since sparsity is desired, Zou and Xue (2006) proposed the following optimization criterion

$$\begin{aligned} \underset{U,V\in\mathbb{R}^{p\times D}}{\textbf{minimize:}} \quad & \frac{1}{n} \left\|(\mathbf{I}_p - UV^T)\mathbf{X}^T\right\|_{\mathrm{F}}^2 + \lambda_1 \left\|V\right\|_2^2 + \lambda_2 \left\|V\right\|_1 \\ \textbf{subject to:} \quad & U^T U = \mathbf{I}_D \ , \end{aligned}$$

and to solve this optimization problem, they use an alternating optimization algorithm. Fixing U, solving V is equivalent to solving $\min_V \frac{1}{n} \left\|\mathbf{X}U - \mathbf{X}V\right\|_{\mathrm{F}}^2 + \lambda_1 \left\|V\right\|_2 + \lambda_2 \left\|V\right\|_1$, which can be broken down to D elastic net regression problems. Fixing V, solving U is equivalent to $\min_{U^TU=\mathbf{I}_D} -\mathbf{tr}\left(V^TX^TXU\right)$, which is an SVD problem. Updating U and V alternatively until convergence, one will eventually get $\hat{V}$, whose column span $\hat{\mathcal{V}} = \mathbf{colspan}\langle\hat{\mathbf{V}}\rangle$ is an estimator of leading principal space $\mathcal{V}$.

The algorithm is efficient because it separates the sparsity penalty from eigen-solver. In SPCA-reg, one of the two steps is a regression and the other is an eigen-solver, and sparsity is only applied to the regression problem. This "separation" is a common feature of efficient SPCA algorithms.

In order to get a common support, we only need to change the penalty and optimize the following instead

$$\begin{aligned}\underset{U,V\in\mathbb{R}^{p\times D}}{\mathbf{minimize:}} \quad & \frac{1}{n}\left\|(\mathbf{I}_D - UV^T)\mathbf{X}^T\right\|_{\mathrm{F}}^2 + \lambda\left\|V\right\|_{1,2} \\ \mathbf{subject\ to:} \quad & U^TU = \mathbf{I}_D\ .\end{aligned}$$

This can be solved using Algorithm 1

Algorithm 1 GSPCA via regression (GSPCA-Reg)

```
1: Initialize $V^{(0)} \in \mathbb{R}^{p\times D}$.
2: $S = \mathbf{X}^T\mathbf{X}/n$.
3: For $k = 0, 1, 2, ...,$, repeat the following until convergence:
4: **Normalization:** Let $Z^{(k+1)} = SV^{(k)}$, and suppose SVD of $Z^{(k+1)} = \mathbf{LSR}$. Then update $U$ by:
       $U^{(k+1)} = \mathbf{LR}$
5: **Group Lasso:** Let $Y^{(k+1)} = \mathbf{X}U^{(k+1)}$; Get $V^{(k+1)}$ by solving:
   (2.1)    **minimize:** $\frac{1}{n}\left\|Y^{(k+1)} - \mathbf{X}V\right\|_{\mathrm{F}}^2 + \lambda\left\|V\right\|_{1,2}$
   which can be solved by block coordinate descent.
6: $\hat{\mathbf{V}} = V^{(k)}$, $\hat{\mathcal{V}} = \mathbf{colspan}\langle V^{(k)}\rangle$.
```

2.2.2 SPCA with soft-thresholded power method (SPCA-Power)

To get the top PC, Witten et al. (2009) proposed the following optimization criterion

$$\begin{aligned}\mathbf{maximize}_{u\in\mathbb{R}^n, v\in\mathbb{R}^p}: \quad & u^T\mathbf{X}v \\ \mathbf{subject\ to}: \quad & \|u\|_2 \leq 1, \|v\|_2 \leq 1, \|v\|_1 \leq \tau\ .\end{aligned}$$

For any fixed v, the optimal u is $u = \mathbf{X}v / \left\|\mathbf{X}v\right\|_2$; if we take in the optimal u, then the above criterion becomes ScotLass criterion. However, this criterion naturally suggests an alternate optimization scheme. When fixing u, optimizing over v is to find the minimal soft-threshold c on $\mathbf{X}^T u$ such that

$$\left\| \frac{\mathcal{T}(\mathbf{X}^T u, c)}{\left\|(\mathcal{T}(\mathbf{X}^T u, c)\right\|_2} \right\|_1 \leq \tau \ .$$

The left hand side is monotonic in c, so finding c is a one-dimensional nonlinear problem, and is easy to compute.

To get subsequent loadings, Witten et al. (2009) proposed an sequential algorithm:

$$\begin{aligned} \underset{u_k \in \mathbb{R}^n, v_k \in \mathbb{R}^p}{\textbf{maximize:}} \quad & u_k^T \mathbf{X} v_k \\ \textbf{subject to:} \quad & \left\|u_k\right\|_2 \leq 1, u_k \perp u_1, ..., u_{k-1} \\ & \left\|v_k\right\|_2 \leq 1, \left\|v_k\right\|_1 \leq \tau \ . \end{aligned}$$

In order to get a common support, we can optimize over all eigenvectors together, for instance

$$\begin{aligned} \underset{U \in \mathbb{R}^n, V \in \mathbb{R}^p}{\textbf{maximize:}} \quad & \textbf{tr}\left(U^T \mathbf{X} V\right) \\ \textbf{subject to:} \quad & U^T U = \mathbf{I}_D, \left\|V\right\|_{\mathrm{F}}^2 \leq D, \left\|V\right\|_{1,2} \leq \tau \ . \end{aligned}$$

Note that $\left\|V\right\|_{\mathrm{F}}^2 \leq D$ is quite a big relaxation of the orthonormal restriction $V^T V = \mathbf{I}_D$. With this relaxation, optimizing V with U fixed is easy to solve according to the following lemma.

Lemma II.2. *For $W \in \mathbb{R}^{p \times D}$, let $w = (w_1, ..., w_p)$ such that $w_j = \left\|W_{j\cdot}\right\|_2$. To solve*

$$\max_{\substack{\left\|V\right\|_{\mathrm{F}}^2 \leq D \\ \left\|V\right\|_{1,2} \leq \tau}} \textbf{tr}\left(W^T V\right) \ , \tag{2.2}$$

it is sufficient to find the soft threshold c such that

$$\left\| \frac{\mathcal{T}(w, c)}{\left\|\mathcal{T}(w, c)\right\|_2} \right\|_1 = \frac{\tau}{\sqrt{D}} \ ,$$

and then $\hat{V}$ *can be obtained by*

$$\hat{V}_{j\cdot} = \frac{W_{j\cdot}}{w_j} \times \frac{(w_j - c)_+}{\|(w_j - c)_+\|_2} \times \sqrt{D}.$$

Optimizing U with V fixed is an SVD problem. In summary we get algorithm 2 which is a orthogonal iteration algorithm with an added soft thresholding.

Algorithm 2 GSPCA through soft-thresholded power method (GSPCA-Power)

```
Initialize $V^{(0)} \in \mathbb{R}^{p \times D}$.
For $k = 0, 1, 2, ...,$, repeat the following until convergence:
**SVD:** Conduct SVD on $\mathbf{X}V^{(k)} = \mathbf{LSR}$. Let
    $$U^{(k)} = \mathbf{LR}$$
**Soft thresholding:** Let $W = \mathbf{X}^T U^{(k+1)}$, $w = (w_1, ..., w_p)$ where $w_j = \|W_{j\cdot}\|_2$. Solve for $c$
    $$\left\| \frac{\mathcal{T}(w, c)}{\|(\mathcal{T}w, c)\|_2} \right\|_1 = \frac{\tau}{\sqrt{D}}$$
    Update $V^{(k+1)}$ so that $V^{(k+1)}_{j\cdot} = \frac{W_{j\cdot}}{w_j} \times \frac{(w_j - c)_+}{\|(w_j - c)_+\|_2} \times \sqrt{D}$
$\hat{\mathbf{V}} = V^{(k)}$, $\hat{\mathcal{V}} = \mathbf{colspan}\langle V^{(k)} \rangle$.
```

2.2.3 SPCA via semidefinite programming (SPCA-SDP)

A clever SDP relaxation of SPCA was proposed by d'Aspremont et al. (2005). To get the leading eigenvector, instead of optimizing over eigenvector v, they optimize over $F = vv^T$. The feasible set of F consists of rank-1 projection matrices. This is a non-convex set, so the optimization over it is hard to compute. A convex relaxation $\mathbf{tr}\,(F) = 1$ is then used instead.

To estimate subsequent PC's, d'Aspremont et al. (2005) propose a sequential style algorithm. If the goal is to estimate the leading principal space, then we can also solve all eigenvectors together. Before relaxation and sparse penalization, the following optimization gives leading eigen-space

$$\begin{aligned} \textbf{maximize: :} \quad & \mathbf{tr}\left(S_n V V^T\right) \\ \textbf{subject to:} \quad & VV^T \text{ is a rank } D \text{ projection matrix .} \end{aligned}$$

Relax the feasible set of $F = VV^T$ and we get

$$\begin{aligned} \textbf{maximize:} \quad & \mathbf{tr}\,(S_n F) \\ \textbf{subject to:} \quad & \|F\|_* \leq D, \|F\|_{\text{op}} \leq 1 \ . \end{aligned}$$

Since $F = VV^T$, where V is supposed to have a few non-zero rows, F is thus sparse at least. We can add a sparse penalty

$$\begin{aligned} \textbf{maximize:} \quad & \mathbf{tr}\,(S_n F) + \lambda \|F\|_1 \\ \textbf{subject to:} \quad & \|F\|_* \leq D, \|F\|_{\text{op}} \leq 1 \ , \end{aligned}$$

which can be solved using ADMM similar to the SCCA algorithm proposed by Gao et al. (2017). Specifically, the augmented Lagrangian is

$$\begin{aligned} \mathcal{L}_\eta(F, G, H) = -\mathbf{tr}\,(S_n F) + \lambda \|F\|_1 + \infty \mathbf{1}\{\|G\|_{\text{op}} > 1\} + \infty \mathbf{1}\{\|G\|_* > D\} \\ + \langle H, F - G\rangle + \frac{\eta}{2} \|F - G\|_{\text{F}}^2 \ . \end{aligned}$$

The ADMM scheme is to iteratively update F, G, H

$$F^{(t+1)} = \arg\min \mathcal{L}_\eta(F^{(t)}, G^{(t)}, H^{(t)}) \ , \tag{2.3}$$

$$G^{(t+1)} = \arg\min \mathcal{L}_\eta(F^{(t+1)}, G^{(t)}, H^{(t)}) \ , \tag{2.4}$$

$$H^{(t+1)} = H^{(t)} + \eta(F^{(t+1)} - G^{(t+1)}) \ . \tag{2.5}$$

One of the key updating steps (2.3) is equivalent to solving

$$F^{(t+1)} = \arg\min \frac{\eta}{2} \left\| F - G^{(t)} + \frac{1}{\eta}(H^{(t)} - S_n) \right\|_{\text{F}}^2 + \lambda \|F\|_1$$

whose solution is soft thresholding $F^{(t+1)} = \mathcal{T}\left(G^{(t)} - (H^{(t)} - S_n)/\eta, \lambda/\eta\right)$. Another updating step (2.4) is equivalent to

$$G^{(t+1)} = \mathop{\arg\min}_{\substack{\|G\|_* \leq D \\ \|G\|_{\text{op}} \leq 1}} \left\| G - F^{(t+1)} - H^{(t)}/\eta \right\|_{\text{F}}^2 \ .$$

Let the SVD of $F^{(t+1)} + H^{(t)}/\eta$ be $\mathbf{LSR}$, where $\mathbf{S} = \mathbf{diag}(\mathbf{s}_1, ..., \mathbf{s}_p)$. Then the solution $G^{(t+1)} = \mathbf{LS}'\mathbf{R}$, where $\mathbf{S}' = \mathbf{diag}(\mathbf{s}'_1, ..., \mathbf{s}'_p)$, is obtained by“capped soft thresholding”, i.e. finding c so that $\mathbf{s}'_j = 1 \wedge (\mathbf{s}_j - c)$ satisfies $\sum_{j=1}^{p} \mathbf{s}'_j = D$. The algorithm is summarized in Algorithm 3.

Algorithm 3 SPCA through SDP relaxation (SPCA-SDP)

```
Initialize $F^{(0)}, G^{(0)} \in \mathbb{R}^{p\times p}$, $H^{(0)} = \mathbf{0}_{p\times p}$.
For $k = 0, 1, 2, ...,$, repeat the following until convergence:
$W^{(t)} = G^{(t)} - (H^{(t)} - S_n)/\eta$
$F^{(t+1)} = \mathcal{T}\left(W^{(t)}, \lambda/\eta\right)$.
SVD: $F^{(t+1)} + H^{(t)}/\eta = \mathbf{LSR}$.
$G^{(t+1)} = \mathbf{LS}'\mathbf{R}$, where $\mathbf{S}'$ is capped soft thresholding of $\mathbf{S}$, i.e. find $c$ so that $\mathbf{s}'_j = 1 \wedge (\mathbf{s}_j - c)$
    satisfies $\sum_j \mathbf{s}'_j = D$.
$H^{(t+1)} = H^{(t)} + \eta(F^{(t+1)} - G^{(t+1)})$.
$\hat{F} = (\hat{F} + \hat{F}^T)/2$ (can be skipped because $F^{(t)}$ will already be symmetric).
$\hat{V}$ is the leading $D$ eigenvectors of $\hat{F}$.
```

Note that the estimated vectors do not have common support. In order to get common support, we can modify the Algorithm 3. When updating F, we soft threshold $W^{(t)} = G^{(t)} - (H^{(t)} - S_n)/\eta$ by rows/columns. More specifically, let the norm of rows to be $w_j = \left\| W^{(t)}_{j\cdot} \right\|_2$, and specify a threshold τ. Apply soft thresholding on the rows $\tilde{W}^{(t)}_{j\cdot} = W^{(t)}_{j\cdot} \frac{(w_j - \tau)_+}{w_j}$. The resulting matrix is row sparse but not symmetric. If we symmetrize it by $\tilde{W}^{(t)} \leftarrow (\tilde{W}^{(t)} + \tilde{W}^{(t)T})/2$, the result is no longer sparse row-wise or column-wise. A practical fix is to let $F^{(t+1)}_{jk} = \tilde{W}^{(t)}_{jk}$ if $w_j, w_k > \tau$, or 0 otherwise. Everything else is kept the same.

Algorithm 4 GSPCA through SDP relaxation (GSPCA-SDP)

```
Initialize $F^{(0)}, G^{(0)} \in \mathbb{R}^{p\times p}$, $H^{(0)} = \mathbf{0}_{p\times p}$.
For $k = 0, 1, 2, ...,$, repeat the following until convergence:
$W^{(t)} = G^{(t)} - (H^{(t)} - S_n)/\eta$, and let $w_j = \left\| W^{(t)}_{j\cdot} \right\|_2$.
Find the set $\hat{I} = \{j : w_j > \tau\}$; do soft thresholding by rows, $\tilde{W}^{(t)}_{j\cdot} = \frac{(w_j - \tau)_+}{w_j} W^{(t)}_{j\cdot}$.
$F^{(t+1)}_{jk} = (\tilde{W}^{(t)}_{jk} + \tilde{W}^{(t)}_{kj})/2$ if $j, k \in \hat{I}$, or otherwise 0.
SVD: $F^{(t+1)} + H^{(t)}/\eta = \mathbf{LSR}$.
$G^{(t+1)} = \mathbf{LS}'\mathbf{R}$, where $\mathbf{S}'$ is capped soft thresholding of $\mathbf{S}$, $\mathbf{s}'_j \le 1$, $\sum_j \mathbf{s}'_j = D$.
$H^{(t+1)} = H^{(t)} + \eta(F^{(t+1)} - G^{(t+1)})$.
$\hat{F} = (\hat{F} + \hat{F}^T)/2$ (can be skipped because $F^{(t)}$ is already symmetric).
$\hat{V}$ is the leading $D$ eigenvectors of $\hat{F}$.
```

2.3 SPCA based on thresholding

2.3.1 SPCA with diagonal thresholding (SPCA-DT)

Due to the inconsistency of regular PCA in high dimensions, a naive approach is to first screen the variables, and run PCA on only the few "important" variables. The screening criterion is inevitably dependent on the model. Recall that our model assumption is a fixed-design spiked model, i.e.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T + \boldsymbol{\epsilon}\ . \tag{2.6}$$

We measure the signal level of the j-th coordinate by $\|\mathbf{V}_{j\cdot}\|_2$, and we want to screen out those coordinates whose signals are zero or very small. For simplicity we assume $\sum_i \mathbf{u}_{id} = 0$ for all d; also $\boldsymbol{\epsilon}$ has i.i.d. rows of mean 0 Gaussian vectors. Thus $\mathbf{1}_n^T\mathbf{X}/n \overset{a.s.}{\to} 0$, so $S = \mathbf{X}^T\mathbf{X}/n$ can be used as sample covariance. The j-th diagonal element of S is given by $S_{jj} = \|\mathbf{U}\mathbf{S}\mathbf{V}_{j\cdot} + \boldsymbol{\epsilon}_{\cdot j}\|_2^2$. If the covariance of noise Σ_e is small compared to $\mathbf{S}$, then $S_{jj} \approx \|\mathbf{U}\mathbf{S}\mathbf{V}_{j\cdot}\|_2^2 = \mathbf{V}_{j\cdot}^T\mathbf{S}^2\mathbf{V}_{j\cdot}$. If $\|\mathbf{V}_{j\cdot}\|_2$ is close to 0, then S_{jj} is also close to 0. Thus we can use the diagonal elements of $\mathbf{X}^T\mathbf{X}/n$ as a thresholding criterion to screen the coordinates. This is called diagonal thresholding, and was originally proposed by Johnstone and Lu (2009). The algorithm is summarized in Algorithm 5.

Algorithm 5 SPCA with diagonal thresholding (SPCA-DT)

```
Given dimension $D$ and threshold $\gamma_1$.
$S = \mathbf{X}^T\mathbf{X}/n$.
$\hat{I}(\gamma_1) = \{j : S_{jj} > \gamma_1\}$.
$\hat{V}^{DT} = (\hat{v}_1^{DT}, ..., \hat{v}_D^{DT})$ are leading eigenvectors of $\begin{pmatrix} S_{\hat{I}\hat{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$.
```

The threshold is typically chosen to be of order $\gamma_1 = c_1\sqrt{\frac{\log p}{n}}$ when working with spiked-covariance model. This choice is sufficient to reliably exclude coordinates of zero or small signals. Note that when noise is small, the j-th diagonal $S_{jj} \approx \mathbf{V}_{j\cdot}^T \mathbf{S}^2 \mathbf{V}_{j\cdot} \asymp \|\mathbf{V}_{j\cdot}\|_2^2$ if $\sigma_1, ..., \sigma_D$ are of constant order. Thus, a rough consequence of such choice is that only those coordinates whose signal levels $\|\mathbf{V}_{j\cdot}\|_2$ are of larger order than $O(\log(p/n)^{1/4})$ can be reliably included. This is not good enough and some improvement is needed. We need to point out that this choice works if the noises are i.i.d., i.e. $\Sigma_e = \sigma_e^2 \mathbf{I}_p$; for general Σ_e, it is not justified.

It is important to conduct thresholding before solving for eigen-system; thresholding the sample eigenvectors generally does not work because the perturbation error increases too fast as p increases. Figure 2.1 shows the probability of getting the correct sparsity using pre-screening and ad hoc thresholding. The data are generated from $\mathbf{X} = uv^T + \boldsymbol{\epsilon}$, where $\sum_i u_i = 0$, $\|u\|_2 = 1$, $v = (1, 0, 0, ..., 0)$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times p}$ has i.i.d. standard Gaussian entries. The correct sparse set is the first coordinate. Using pre-screening, we need $\|\mathbf{X}_{\cdot 1}\|_2 \geq \max_{j>1} \|\mathbf{X}_{\cdot j}\|_2^2$; using ad hoc thresholding, we first calculate top right singular vector $\hat{v}$, and require $|\hat{v}_1| > \max_{j>1} |\hat{v}_j|$. We let $n = 50, 100, ..., 400$, and $p = n/2$ or $p = 2n$, and check the proportion of 2000 independent experiments, for which the true signal coordinate is successfully separated from the noise coordinates. The results show that pre-screening works well even when $p > n$; in fact p can be of exponential order to n when using Gaussian noise. In contrast, ad hoc thresholding does not work. We essentially need $\hat{v}$ to be consistent, which will not occur in high dimensional case.

2.3.2 SPCA with augmented thresholding (SPCA-AT)

It was shown in both Birnbaum et al. (2013) (Theorem 4.1) and Ma (2013) (remark after Theorem 3.3) that although principal subspace estimated by SPCA-DT can be consistent, it is not rate optimal. The common argument is: in order to achieve optimal error rate, we need to be able to differentiate a coordinate whose “signal” is

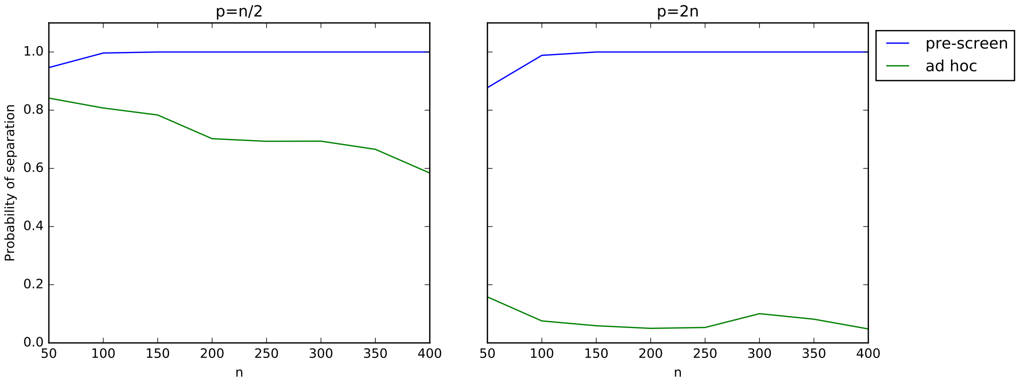


Figure 2.1: Probability of estimating the correct sparse set

of order $(\log p/n)^{1/2}$ from those coordinates of zero or "small" signals, but SPCA-DT can only differentiate signal levels of order $(\log p/n)^{1/4}$. Technical differences exist regarding how sparsity or signal level is defined, and what is considered as "small" signal, but the main proof techniques are always similar.

It is widely stated that DT is not ideal because only diagonal information is used, so a common remedy is to run another screening that utilizes off-diagonal information, in order to potentially includes more coordinates, especially those whose signal levels are between the order of $(\log p/n)^{1/2}$ and $(\log p/n)^{1/4}$.

To motivate this improved coordinate screening, consider $S\hat{\mathbf{V}}$ for some estimator $\hat{\mathbf{V}}$ of $\mathbf{V}$. If the noise is small, and $\hat{\mathbf{V}}$ is a good estimate for $\mathbf{V}$, then $S\hat{\mathbf{V}} \approx \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$. Then $\left\|[S\hat{\mathbf{V}}]_{j\cdot}\right\|_2$ is close to 0 as long as $\left\|\hat{\mathbf{V}}_{j\cdot}\right\|_2$ is close to 0. Thus, if we have a good estimated sparse set $\hat{I}$ to start with, we can let $\hat{\mathbf{V}}$ be the PC loading vectors restricted to $\hat{I}$, and use $\left\|[S\hat{\mathbf{V}}]_{j\cdot}\right\|_2$ as the criterion to do one more round of thresholding. This method is proposed by Birnbaum et al. (2013) as augmented SPCA (ASPCA); we denote it as SPCA-AT (augmented thresholding). The algorithm is described in Algorithm 6.

Threshold of level $\gamma_2 = c_2\sqrt{\frac{\log p}{n}}$ is recommended to ensure that zero coordinates are correctly excluded. Intuitively, for those non-zero coordinates, $W_j \approx \mathbf{\Lambda}\mathbf{V}_{j\cdot}$,

Algorithm 6 SPCA with augmented thresholding (SPCA-AT)

```
Given dimension $D$ and threshold parameter $\gamma_1, \gamma_2$.
$S = \mathbf{X}^T\mathbf{X}/n$.
**Diagonal thresholding:** Using SPCA-DT to obtain the initial subset:
    $$\hat{I} = \hat{I}(\gamma_1); \quad \hat{V} = \hat{V}^{DT}\ .$$
$W = S\hat{V} \in \mathbb{R}^{p\times D}$ $w_j = \|W_{j\,\cdot}\|_2$.
**Augmented thresholding:** $\tilde{I}(\hat{I}, \gamma_2) = \{j \notin \hat{I} : w_j > \gamma_2\} \cup \hat{I}$.
$\hat{V}^{AT} = (\hat{v}_1^{AT}, ..., \hat{v}_D^{AT})$ are leading eigenvectors of $\begin{pmatrix} S_{\tilde{I}\tilde{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$.
```

so $\|W_{j\,\cdot}\|_2 \asymp \|\mathbf{V}_{j\,\cdot}\|_2$ if $\sigma_1, ..., \sigma_D$ are of constant order. Therefore the threshold can differentiate those coordinates whose signals are of order $\sqrt{\frac{\log p}{n}}$. This choice of threshold applies to i.i.d. noise and need to be modified for general Σ_e.

2.4 Statistical property of SPCA-DT and SPCA-AT

2.4.1 Distance between linear subspaces

Our goal is to estimate the leading principal space $\mathcal{V} = \mathbf{colspan}\langle\mathbf{V}\rangle \subseteq \mathbb{R}^p$ with a sparse estimator $\hat{\mathcal{V}} = \mathbf{colspan}\langle\hat{\mathbf{V}}\rangle$ that only use a small subset of variables, so an immediate question is how to measure the difference between two subspaces.

Suppose that we have two sets of orthonormal bases $\mathbf{V}, \hat{\mathbf{V}} \in \mathbb{R}^{p\times D}$. When $D = 1$, the subspace decides the basis vector $\mathbf{v}_1, \hat{\mathbf{v}}_1$ uniquely up to sign ($\mathbf{v}_1$ and $-\mathbf{v}_1$ are both basis of $\mathcal{V}$). One measurement that makes sense for $D = 1$ is then

$$\mathbf{Dist}_{vec}(\mathcal{V}, \hat{\mathcal{V}}) = \min\{\|\mathbf{v}_1 - \hat{\mathbf{v}}_1\|_2, \|\mathbf{v}_1 + \hat{\mathbf{v}}_1\|_2\}\ .$$

When $D > 1$ however orthonormal basis is only unique under orthonormal transformation, i.e. if $\mathbf{V}$ is an orthonormal basis of $\mathcal{V}$, $\mathbf{VO}$ is also an orthonormal basis, for any $\mathbf{O} \in \mathbb{R}^{D\times D}$ s.t. $\mathbf{O}^T\mathbf{O} = \mathbf{I}_D$. Thus, difference between two specific orthonormal bases does not reflect the difference between the spaces. Therefore $\mathbf{Dist}_{vec}$ does not make sense when $D > 1$.

A good distance should be invariant by orthornormal transformation. It can be based on so called **principal angles** or **canonical angels**. The principal angles $(\theta_1, ..., \theta_D) \in [0, \pi/2]$ are recursively defined as

$$\cos(\theta_d) = \frac{|u_d^T v_d|}{\|u_d\|_2 \, \|v_d\|_2} \ ,$$

where

$$u_d, v_d = \arg\max_{u \in \mathcal{V}, v \in \hat{\mathcal{V}}} \frac{|u^T v|}{\|u\|_2 \, \|v\|_2} \ ,$$

subject to

$$u_d \perp u_1, ..., u_{d-1}, \ v_d \perp v_1, ..., v_{d-1} \ .$$

A more convenient but less intuitive definition of the same distance is that $(\cos(\theta_1), ..., \cos(\theta_D))$ are singular values of $\mathbf{V}^T\hat{\mathbf{V}}$. In addition, we can define $\sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) = \mathbf{diag}(\sin(\theta_1), ..., \sin(\theta_D))$. Generally, if $\sin\Theta(\mathcal{V}, \hat{\mathcal{V}})$ is large, then the "difference" between $\mathcal{V}$ and $\hat{\mathcal{V}}$ is large.

There are several distance measurements. One distance measures "average angle" in some sense, i.e.,

$$\mathbf{Dist}_{ave}(\mathcal{V}, \hat{\mathcal{V}})^2 = \frac{1}{D} \left\| \sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) \right\|_{\mathrm{F}}^2 = \frac{1}{D} \sum_{d=1}^{D} \sin(\theta_d)^2 \ .$$

This distance can also be derived from either difference between projection matrices or inner product of the two bases. Note that projection matrix of a subspace is also uniquely defined, $P_{\mathcal{V}} = \mathbf{V}\mathbf{V}^T$, $P_{\hat{\mathcal{V}}} = \hat{\mathbf{V}}\hat{\mathbf{V}}^T$. We have

$$\begin{aligned}
\|P_{\mathcal{V}} - P_{\hat{\mathcal{V}}}\|_{\mathrm{F}}^2 &= \left\| \mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T \right\|_{\mathrm{F}}^2 \\
&= \mathbf{tr}\left(\mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T\right) + \mathbf{tr}\left(\hat{\mathbf{V}}\hat{\mathbf{V}}^T\hat{\mathbf{V}}\hat{\mathbf{V}}^T\right) - 2\mathbf{tr}\left(\mathbf{V}\mathbf{V}^T\hat{\mathbf{V}}\hat{\mathbf{V}}^T\right) \\
&= 2D - 2\left\| \mathbf{V}^T\hat{\mathbf{V}} \right\|_{\mathrm{F}}^2 \\
&= 2D - 2\sum_{d=1}^{D} \cos(\theta_d)^2 = 2\left\| \sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) \right\|_{\mathrm{F}}^2 \ .
\end{aligned}$$

Another distance measures "maximal angle",

$$\mathbf{Dist}_{max}(\mathcal{V}, \hat{\mathcal{V}}) = \left\| \sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) \right\|_{\mathrm{op}}^2 = \sin(\theta_D) \ .$$

The maximal angle can also be derived from the inner product of two bases. We have

$$\cos(\theta_D) = \sigma_{\min}(\mathbf{V}^T\hat{\mathbf{V}}) \ .$$

Thus we have the following proposition that summarizes the relationships between principal angles, inner product and projection matrices.

Proposition II.3. *For two linear subspaces $\mathcal{V}$, $\hat{\mathcal{V}}$, let $\mathbf{V}$ and $\hat{\mathbf{V}}$ be any orthonormal bases of these two subspaces. Then,*

1.

$$\mathbf{Dist}_{ave}(\mathcal{V}, \hat{\mathcal{V}}) = \frac{1}{\sqrt{D}} \left\| \sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) \right\|_{\mathrm{F}} = \frac{1}{\sqrt{2D}} \left\| P_{\mathcal{V}} - P_{\hat{\mathcal{V}}} \right\|_{\mathrm{F}} = \sqrt{1 - \left\| \mathbf{V}^T\hat{\mathbf{V}} \right\|_{\mathrm{F}}^2 / D} \ ;$$

2.

$$\mathbf{Dist}_{max}(\mathcal{V}, \hat{\mathcal{V}}) = \left\| \sin\Theta(\mathcal{V}, \hat{\mathcal{V}}) \right\|_{\mathrm{op}} = \sqrt{1 - \sigma_{\min}(\mathbf{V}^T\hat{\mathbf{V}})^2} \ ;$$

3. When $D = 1$, $\mathbf{Dist}_{ave} = \mathbf{Dist}_{max}$.

For convenience, sometimes we use $\mathbf{Dist}_{vec}(\mathbf{v}, \hat{\mathbf{v}}), \mathbf{Dist}_{ave}(\mathbf{V}, \hat{\mathbf{V}}), \mathbf{Dist}_{max}(\mathbf{V}, \hat{\mathbf{V}})$ instead of $\mathbf{Dist}_{vec}(\mathcal{V}, \hat{\mathcal{V}}), \mathbf{Dist}_{ave}(\mathcal{V}, \hat{\mathcal{V}}), \mathbf{Dist}_{max}(\mathcal{V}, \hat{\mathcal{V}})$.

2.4.2 Single-spike and i.i.d. noise, $D = 1$, $\Sigma_e = \sigma_e^2 \mathbf{I}_p$

Here we first derive non-asymptotic statistical properties for the model with a single-spike and i.i.d. noises. We can simplify the notations

$$\mathbf{X} = \sigma \mathbf{u}\mathbf{v}^T + \boldsymbol{\epsilon} \ , \tag{2.7}$$

where $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^p$ are two fixed vectors, s.t. $\|\mathbf{u}\|_2 = \sqrt{n}$, $\|\mathbf{v}\|_2 = 1$, and $\boldsymbol{\epsilon}_{ij} \sim \mathbf{Norm}(0, \sigma_e^2)$.

Assume that $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_p)$ is sparse in ℓ_0 sense, and let the size of $I_0 = \{j : \mathbf{v}_j \neq 0\}$ be $|I_0| = s$. We want to screen out all coordinates whose signals $\mathbf{v}_k$ are

0, while keeping all the coordinates whose signals are large, i.e. $|\mathbf{v}_k| > \kappa$. Here κ indicates the lower bound of the signal level in order to be differentiable from noise, and we want it to be as small as possible.

Since $D = 1$, we aim to bound $\mathbf{Dist}_{vec}(\mathcal{V}, \hat{\mathcal{V}})$. For simplicity, we often write $\|\hat{\mathbf{v}} - \mathbf{v}\|_2$ directly, with the caveat that $\pm\hat{\mathbf{v}}$ are usually not differentiable, and we assume $\hat{\mathbf{v}}$ to be the one closer to $\mathbf{v}$ (angle smaller than $\pi/2$). Recall that the error measure can be easily translated to the general distance measures; as explained previously, when $D = 1$, $\mathbf{Dist}_{max} = \mathbf{Dist}_{ave} = \sqrt{1 - \cos\theta^2} \leq \sqrt{2(1-\cos\theta)} = \mathbf{Dist}_{vec} \leq \sqrt{2}\mathbf{Dist}_{ave}$.

The basic idea of the proof is to find two small constants $\kappa_1 > \kappa_2 > 0$ and the corresponding subsets $I_1 = \{j : |\mathbf{v}_j| > \kappa_1\}$, $I_2 = \{j : |\mathbf{v}_j| > \kappa_2\}$, so that with high probability, the diagonal thresholding set $\hat{I}$ satisfies $I_1 \subseteq \hat{I} \subseteq I_0$ and the augmented thresholding set $\tilde{I}$ satisfies $I_2 \subseteq \tilde{I} \subseteq I_0$. We will see that κ_1 determines the "bias" of DT estimator $\hat{v}^{DT}$ while κ_2 determines the "bias" of AT estimator $\hat{v}^{AT}$. The size of I_0, s, determines the "variance" of the estimator.

Theorem II.4. *Assume the single-spike model with i.i.d. noises as in (2.7). Suppose that there exists a constant* $0 < \delta < 1$ *and* $t_1, t_2, t_3 > 0$ *such that* $t_0 = \sqrt{\frac{\log p - \log\delta}{n}} \in (0, 1/2)$, *and*

$$\kappa_1 = \frac{\sigma_e}{\sigma}\sqrt{4t_0 + 12t_0^2}\ ,$$

$$\kappa_2 = \kappa_1 \wedge \kappa_2',\ \kappa_2' = \frac{\sigma + \sigma_e + \sigma_e\frac{\sqrt{s}+t_1}{\sqrt{n}}}{\sigma\sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e\frac{3\sqrt{s}+2t_1+t_2}{\sqrt{n}}}\frac{\sigma_e}{\sigma}\frac{\sqrt{2\log p + 2s\log s} + t_3}{\sqrt{n}}$$

satisfies $\kappa_2 > 0$, $\sqrt{s}\kappa_1 < 1$. *Denote* $I_1 = \{j : |\mathbf{v}_j| > \kappa_1\}$, $I_2 = \{j : |\mathbf{v}_j| > \kappa_2\}$. *Then using the following two thresholds*

$$\gamma_1 = \sigma_e^2(1 + 2t_0 + 2t_0^2)\ ,$$

$$\gamma_2 = \sigma_e\left(\sigma + \sigma_e + \sigma_e\frac{\sqrt{s}+t_1}{\sqrt{n}}\right)\frac{\sqrt{2\log p + 2s\log s} + t_3}{\sqrt{n}}\ ,$$

we have,

1. *with probability at least $1-\delta$, the DT estimated set $\hat{I}=\hat{I}(\gamma_1)$ satisfies*

$$I_1 \subseteq \hat{I} \subseteq I_0 \ ;$$

2. *With probability at least $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)$ the DT estimated eigenvector $\hat{\mathbf{v}}^{DT}$ satisfies*

$$\left\|\hat{\mathbf{v}}^{DT}-\mathbf{v}\right\|_2 \le \sqrt{2s}\kappa_1 + \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_1^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_1^2)}\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2-1\right] ,$$

where $\hat{\mathbf{v}}^{DT}$ is the one between $\pm\hat{\mathbf{v}}^{DT}$ that is closer to $\mathbf{v}$;

3. *With probability at least $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)$, the AT estimated set $\tilde{I}=\tilde{I}(\gamma_2,\hat{I}(\gamma_1))$ satisfies*

$$I_2 \subseteq \tilde{I} \subseteq I_0 \ ;$$

4. *With probability at least $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)$ the AT estimated eigenvector $\hat{\mathbf{v}}^{AT}$ satisfies*

$$\left\|\hat{\mathbf{v}}^{AT}-\mathbf{v}\right\|_2 \le \sqrt{2s}\kappa_2 + \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_2^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_2^2)}\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2-1\right] ,$$

where $\hat{\mathbf{v}}^{AT}$ is the one between $\pm\hat{\mathbf{v}}^{AT}$ that is closer to $\mathbf{v}$.

Proof. 1. **Preparation:**

Define the following random events

$$\mathcal{G}_1 : \max_{k\in I_0^c} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} < \sigma_e^2(1+2t_0+2t_0^2) ,$$

$$\mathcal{G}_2 : \min_{k \in I_1} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} > \sigma_e^2(1 + 2t_0 + 2t_0^2) \ ,$$

$$\mathcal{G}_3 : \sigma_{\max}(\boldsymbol{\epsilon}_{\cdot I_0}) \leq \sigma_e(\sqrt{n} + \sqrt{s} + t_1) \ ,$$

$$\mathcal{G}_4 : \sigma_{\min}(\boldsymbol{\epsilon}_{\cdot I_0}) \geq \sigma_e(\sqrt{n} - \sqrt{s} - t_1) \ ,$$

$$\mathcal{G}_5 : \left\| \frac{\mathbf{u}}{\sqrt{n}}^T \boldsymbol{\epsilon}_{\cdot I_0} \right\|_2 \leq \sigma_e(\sqrt{s} + t_2) \ ,$$

$$\mathcal{G}_6 : \max_{k \in I_1^c, I_1 \subseteq I' \subseteq I_0, k \notin I'} \left\| \boldsymbol{\epsilon}_{\cdot k}^T \hat{\mathbf{u}}^{(I')} \right\|_2 \leq \sigma_e(\sqrt{2\log p + 2s\log s} + t_3) \ ,$$

where in $\mathcal{G}_6$, we use $\hat{\sigma}^{(I)}$, $\hat{\mathbf{u}}^{(I)}$ and $\hat{\mathbf{v}}^{(I)}$ to denote the top singular value, the top left singular vector and the right singular vector of $\mathbf{X}_{\cdot I}$ for a subset of indices I, respectively.

2. **Diagonal thresholding set:**

   If $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, then obviously, by setting $\gamma_1 = \sigma_e^2(1 + 2t_0 + 2t_0^2)$, we have the DT subset $I_1 \subseteq \hat{I} \subseteq I_0$. Therefore,

   $$P(I_1 \subseteq \hat{I} \subseteq I_0) \geq 1 - P(\mathcal{G}_1^c) - P(\mathcal{G}_2^c) \ . \tag{2.8}$$

3. **From the DT set to the corresponding estimator:**

   Now we have a random set $\hat{I}$ that is known to be sandwiched between I_1 and I_0, and we want to analyze the error between the eigenvector restricted to $\hat{I}$ and the true eigenvector.

   For any subset I, denote $\hat{\mathbf{v}}^{(I)}$ to be the top right singular vector of $\mathbf{X}_{\cdot I}$. Thus the DT estimator $\hat{\mathbf{v}}^{DT} = \begin{pmatrix} \hat{\mathbf{v}}^{(\hat{I})} \\ \mathbf{0} \end{pmatrix}$. The triangular inequality gives

   $$\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \leq \left\| \hat{\mathbf{v}}^{(\hat{I})} - \frac{\mathbf{v}_{\hat{I}}}{\|\mathbf{v}_{\hat{I}}\|_2} \right\|_2 + \left\| \begin{pmatrix} \mathbf{v}_{\hat{I}} / \|\mathbf{v}_{\hat{I}}\|_2 \\ \mathbf{0} \end{pmatrix} - \mathbf{v} \right\|_2 \ . \tag{2.9}$$

The first term is the variance term and the second term is the bias term. To deal with the bias term, we have

$$
\begin{aligned}
\left\| \begin{pmatrix} \mathbf{v}_{\hat I}/\left\|\mathbf{v}_{\hat I}\right\|_2 \\ \mathbf{0} \end{pmatrix} - \mathbf{v} \right\|_2 &\le \sqrt{\left\| \frac{\mathbf{v}_{\hat I}}{\left\|\mathbf{v}_{\hat I}\right\|_2} - \mathbf{v}_{\hat I} \right\|_2^2 + \left\|\mathbf{v}_{\hat I^c}\right\|_2^2} \\
&= \sqrt{(1-\left\|\mathbf{v}_{\hat I}\right\|_2)^2 + \left\|\mathbf{v}_{\hat I^c}\right\|_2^2} \\
&\le \sqrt{2\left\|\mathbf{v}_{\hat I^c}\right\|_2} \le \sqrt{2}\left\|\mathbf{v}_{I_1}^c\right\|_2 .
\end{aligned}
\tag{2.10}
$$

The variance term can be dealt with using the perturbation bound of top eigenvector. Note that $\hat{\mathbf{v}}_{\hat I}$ is the top eigenvector of

$$
(\sigma \mathbf{u}\mathbf{v}_{\hat I}^T + \boldsymbol{\epsilon}_{\cdot \hat I})^T(\sigma \mathbf{u}\mathbf{v}_{\hat I}^T + \boldsymbol{\epsilon}_{\cdot \hat I})/n = (\sigma^2 \mathbf{v}_{\hat I}\mathbf{v}_{\hat I}^T + \sigma_e^2 \mathbf{I}_{|\hat I|}) + (M_1 + M_1^T + M_2) ,
$$

where

$$
\begin{aligned}
M_1 &= \sigma \mathbf{v}_{\hat I}\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat I}/n , \\
M_2 &= (\boldsymbol{\epsilon}_{\cdot \hat I}^T \boldsymbol{\epsilon}_{\cdot \hat I}^T/n - \sigma_e^2 \mathbf{I}_{|\hat I|}) .
\end{aligned}
$$

We can view $M_1 + M_1^T + M_2$ as the perturbation term; $(\sigma^2 \mathbf{v}_{\hat I}\mathbf{v}_{\hat I}^T + \sigma_e^2 \mathbf{I}_{|\hat I|})$ is the signal term whose leading eigenvector is $\frac{\mathbf{v}_{\hat I}}{\left\|\mathbf{v}_{\hat I}\right\|_2}$, and the top two eigenvalues are $\sigma^2\left\|\mathbf{v}_{\hat I}\right\|_2^2 + \sigma_e^2$ and σ_e^2, respectively. Thus the eigenvalue gap is $\sigma^2\left\|\mathbf{v}_{\hat I}\right\|_2^2 \ge \sigma^2\left\|\mathbf{v}_{I_1}\right\|_2^2$. By Davis-Kahan's inequality (Lemma II.18), we have

$$
\left\| \hat{\mathbf{v}}^{(\hat I)} - \frac{\mathbf{v}_{\hat I}}{\left\|\mathbf{v}_{\hat I}\right\|_2} \right\|_2 \le \frac{2\sqrt{2}\left\|M_1 + M_1^T + M_2\right\|_{\mathrm{op}}}{\sigma^2\left\|\mathbf{v}_{I_1}\right\|_2^2} .
\tag{2.11}
$$

Now, to bound $\left\|M_1 + M_1^T + M_2\right\|_{\mathrm{op}} \le 2\left\|M_1\right\|_{\mathrm{op}} + \left\|M_2\right\|_{\mathrm{op}}$, we use $\mathcal{G}_3$, $\mathcal{G}_4$ and $\mathcal{G}_5$. For M_1, it is of rank 1, so

$$
\left\|M_1\right\|_{\mathrm{op}} = \sigma\left\|\mathbf{v}_{\hat I}\right\|_2 \left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat I}\right\|_2/n .
$$

Note that $\hat I \subseteq I_0$, so $\left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat I}\right\|_2 \le \left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot I_0}\right\|_2$. By $\mathcal{G}_5$,

$$
\left\|M_1\right\|_{\mathrm{op}} \le \sigma\sigma_e\left\|\mathbf{v}_{\hat I}\right\|_2(\sqrt{s} + t_2)/\sqrt{n} .
\tag{2.12}
$$

For M_2, it is sub-block of $(\boldsymbol{\epsilon}^T_{\cdot I_0}\boldsymbol{\epsilon}^T_{\cdot I_0}/n - \sigma_e^2\mathbf{I}_s)$, and the operator norm of the sub-block is smaller than the whole matrix. Thus by $\mathcal{G}_3, \mathcal{G}_4$,

$$\begin{aligned}
\|M_2\|_{\text{op}} &\le \left\|\boldsymbol{\epsilon}^T_{\cdot I_0}\boldsymbol{\epsilon}^T_{\cdot I_0}/n - \sigma_e^2\mathbf{I}_s\right\|_{\text{op}} \\
&\le \sigma_e^2 \max\left\{\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 - 1, 1-\left(1-\sqrt{\frac{s}{n}}-\frac{t_1}{\sqrt{n}}\right)^2\right\} \qquad (2.13)\\
&\le \sigma_e^2\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 - 1\right] .
\end{aligned}$$

Take (2.12) and (2.13) into (2.11), we get

$$\left\|\hat{\mathbf{v}}^{(\hat{I})} - \frac{\mathbf{v}_{\hat{I}}}{\|\mathbf{v}_{\hat{I}}\|_2}\right\|_2 \le \frac{4\sqrt{2}\sigma_e}{\sigma\|\mathbf{v}_{I_1}\|_2}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2\|\mathbf{v}_{I_1}\|_2^2}\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 - 1\right] . \tag{2.14}$$

In summary, take 2.10 and 2.14 into 2.9 and use the fact that $\left\|\mathbf{v}_{I_1^c}\right\|_2 \le \sqrt{s}\kappa_1$, we have that, with probability at least $1-\sum_{j=1,2,3,4,5} P(\mathcal{G}_j^c)$,

$$\begin{aligned}
\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \le \sqrt{2s}\kappa_1 &+ \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_1^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right) \\
&+ \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_1^2)}\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 - 1\right] .
\end{aligned} \tag{2.15}$$

4. **Second thresholded set:**

For any $k \le p$ and $I \subseteq \{1, 2, ..., p\}$, denote

$$\Delta_{kI} = \boldsymbol{\epsilon}^T_{\cdot k}\mathbf{X}_{\cdot I}/n ,$$

so that

$$\Delta_{kI}\hat{\mathbf{v}}^{(I)} = \hat{\sigma}^{(I)}\boldsymbol{\epsilon}^T_{\cdot k}\hat{\mathbf{u}}^{(I)}/n ,$$

$$S_{kI} = \mathbf{X}^T_{\cdot k}\mathbf{X}_{\cdot I}/n = \sigma\mathbf{v}_k(\sigma\mathbf{v}^T_{\cdot I}/n + \mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I}/n) + \Delta_{kI} .$$

For $l \in I_0^c$, we have $S_{l\hat{I}} = \Delta_{l\hat{I}}$, and hence

$$w_l = \left|\Delta_{l\hat{I}}\hat{\mathbf{v}}^{(\hat{I})}\right| = \hat{\sigma}^{(\hat{I})}\left|\boldsymbol{\epsilon}^T_{\cdot k}\hat{\mathbf{u}}^{(\hat{I})}\right|/n . \tag{2.16}$$

For $k \in I_0 \backslash \hat{I}$, we have

$$\begin{aligned} w_k &= \left| \sigma \mathbf{v}_k \left(\sigma \left\| \mathbf{v}_{\hat{I}} \right\|_2 \times \left(\frac{\mathbf{v}_{\hat{I}}}{\left\| \mathbf{v}_{\hat{I}} \right\|} \right)^T \hat{\mathbf{v}}^{(\hat{I})} + \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}} \hat{\mathbf{v}}^{(\hat{I})} / n \right) + \Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})} \right| \\ &\geq \sigma \mathbf{v}_k \left(\sigma \left\| \mathbf{v}_{\hat{I}} \right\|_2 \left| \left(\frac{\mathbf{v}_{\hat{I}}}{\left\| \mathbf{v}_{\hat{I}} \right\|} \right)^T \hat{\mathbf{v}}^{(\hat{I})} \right| - \left\| \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}} / n \right\|_2 \right) - \hat{\sigma}^{(\hat{I})} \left| \boldsymbol{\epsilon}^T_{\cdot k} \hat{\mathbf{u}}^{(\hat{I})} \right| / n \; . \end{aligned} \tag{2.17}$$

Since $I_1 \subseteq \hat{I} \subseteq I_0$,

$$\left\| \mathbf{v}_{\hat{I}} \right\|_2 \geq \left\| \mathbf{v}_{I_1} \right\|_2 \geq \sqrt{1 - s\kappa_1^2} \; , \tag{2.18}$$

and by $\mathcal{G}_3$ and $\mathcal{G}_5$, respectively, we have

$$\sigma_{\max}(\boldsymbol{\epsilon}_{\hat{I}}) \leq \sigma_{\max}(\boldsymbol{\epsilon}_{I_0}) \leq \sigma_e(\sqrt{n} + \sqrt{s} + t_1) \; ,$$

$$\left\| \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}} / \sqrt{n} \right\|_2 \leq \left\| \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot I_0} / \sqrt{n} \right\|_2 \leq \sigma_e(\sqrt{s} + t_2) \; . \tag{2.19}$$

Since $\sigma^{(\hat{I})}$, $\hat{\mathbf{v}}^{(\hat{I})}$ are singular value and vector of $M = \sigma \mathbf{u} \mathbf{v}_{\hat{I}}^T$ plus perturbation $E = \boldsymbol{\epsilon}_{\cdot \hat{I}}$, we can use Weyl's inequality (Lemma II.15) and Wedin's inequality (Lemma II.16) to get

$$\hat{\sigma}^{(\hat{I})} \leq \sigma \sqrt{n} \left\| \mathbf{v}_{\hat{I}} \right\|_2 + \sigma_e(\sqrt{n} + \sqrt{s} + t_1) \; , \tag{2.20}$$

$$\left| \left(\frac{\mathbf{v}_{\hat{I}}}{\left\| \mathbf{v}_{\hat{I}} \right\|} \right)^T \hat{\mathbf{v}}^{(\hat{I})} \right| \geq 1 - 2 \frac{\sigma_e(\sqrt{n} + \sqrt{s} + t_1)}{\sigma \sqrt{n} \left\| \mathbf{v}_{\hat{I}} \right\|_2} \; ,$$

$$\begin{aligned} \left\| \mathbf{v}_{\hat{I}} \right\|_2 \left| \left(\frac{\mathbf{v}_{\hat{I}}}{\left\| \mathbf{v}_{\hat{I}} \right\|} \right)^T \hat{\mathbf{v}}^{(\hat{I})} \right| &\geq \left\| \mathbf{v}_{\hat{I}} \right\|_2 - 2 \frac{\sigma_e(\sqrt{n} + \sqrt{s} + t_1)}{\sigma \sqrt{n}} \\ &\geq \sqrt{1 - s\kappa_1^2} - 2 \frac{\sigma_e(\sqrt{n} + \sqrt{s} + t_1)}{\sigma \sqrt{n}} \; . \end{aligned} \tag{2.21}$$

Finally, since $\hat{I}$ and any $k \notin \hat{I}$ form a pair that satisfies the condition in $\mathcal{G}_6$, we then have that for $l \in I_0^c$ and $k \in I_0 \backslash \hat{I}$,

$$\left| \boldsymbol{\epsilon}^T_{\cdot k} \hat{\mathbf{u}}^{(\hat{I})} \right| , \left| \boldsymbol{\epsilon}^T_{\cdot l} \hat{\mathbf{u}}^{(\hat{I})} \right| \leq \sigma_e(\sqrt{2 \log p + 2s \log s} + t_3) \; . \tag{2.22}$$

Take all these ingredients (2.18)-(2.22) into (2.16) and (2.17), we have

$$w_l \leq \gamma_2, \quad \forall l \in I_0^c, \tag{2.23}$$

$$w_k \geq \sigma \mathbf{v}_k \left(\sigma \sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e \frac{3\sqrt{s} + 2t_1 + t_2}{\sqrt{n}} \right) - \gamma_2, \tag{2.24}$$

$$\gamma_2 = \sigma_e \left(\sigma + \sigma_e + \sigma_e \frac{\sqrt{s} + t_1}{\sqrt{n}} \right) \frac{\sqrt{2 \log p + 2s \log s} + t_3}{\sqrt{n}}. \tag{2.25}$$

Therefore, if we let

$$\kappa_2' := \frac{\sigma + \sigma_e + \sigma_e \frac{\sqrt{s}+t_1}{\sqrt{n}}}{\sigma\sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e \frac{3\sqrt{s}+2t_1+t_2}{\sqrt{n}}} \frac{2\sigma_e}{\sigma} \frac{\sqrt{2 \log p + 2s \log s} + t_3}{\sqrt{n}} ,$$

then $\mathbf{v}_k > \kappa_2'$ implies $k \in \tilde{I}$. Thus, with probability at least $1 - \sum_{j=1}^6 P(\mathcal{G}_j^c)$, $I_2 \subseteq \tilde{I} \subseteq I_0$.

5. **From the second set to the corresponding estimator:**

   We know that $I_2 \subseteq \tilde{I} \subseteq I_0$, where I_2, I_0 are two fixed sets. We can use the same argument as in part 3 to conclude that, with probability at least $1 - \sum_{j=1}^6 P(\mathcal{G}_j^c)$, we have

$$\begin{aligned} \left\| \hat{\mathbf{v}}^{DT} - \mathbf{v} \right\|_2 \leq \sqrt{2s}\kappa_2 &+ \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1 - s\kappa_2^2}} \left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}} \right) \\ &+ \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1 - s\kappa_2^2)} \left[\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}} \right)^2 - 1 \right]. \end{aligned} \tag{2.26}$$

6. **Bound the probabilities:**

   The rest of the proof is to bound the probabilities of the random events $\mathcal{G}_1^c, ..., \mathcal{G}_6^c$. For $\mathcal{G}_1$, note that for $k \in I_0^c$, $\mathbf{X}_{\cdot k} = \boldsymbol{\epsilon}_{\cdot k}$. Marginally $\|\boldsymbol{\epsilon}_{\cdot k}\|_2^2 \sim \sigma_e^2 \chi_n^2$. Thus by Corollary II.11, we have

$$P\left(\frac{\|\boldsymbol{\epsilon}_{\cdot k}\|_2^2}{n} \geq \sigma_e^2(1 + 2t_0 + 2t_0^2) \right) = P\left(\chi_n^2/n > 1 + 2t_0 + 2t_0^2 \right) \leq \exp(-t_0^2 n) ,$$

and hence, by union bound, $P(\mathcal{G}_1^c) \le \sum_{k \in I_0^c} \exp(-t_0^2 n) = (p-s)\delta/p$.

For $\mathcal{G}_2$, note that for $k \in I_1$, $\frac{\mathbf{X}_{\cdot k}}{\sigma_e} \sim \mathbf{Norm}(\mu, \mathbf{I}_n)$, where $\mu = \frac{\mathbf{v}_k \sigma}{\sigma_e}\mathbf{u}$, $\|\mu\|_2 = \frac{\sqrt{n}\sigma\mathbf{v}_k}{\sigma_e}$. Then, by Lemma II.12,

$$P\left(\frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n\sigma_e^2} \le \left(1 + \frac{\sigma^2\mathbf{v}_k^2}{\sigma_e^2}\right) - 2\sqrt{1 + \frac{2\sigma^2\mathbf{v}_k^2}{\sigma_e^2}} t_0\right) \le \exp(-t_0^2 n) \ .$$

Consider the function $h(x) = (1+x) - 2t_0\sqrt{1+2x}$, and notice that $h(x)$ is monotonically increasing on $x > 2t_0^2 - 0.5$. By our assumption, $2t_0^2 - 0.5 < 0$, so $h(x)$ is increasing on $x > 0$. Now since $k \in I_1$, by construction of I_1, we have $\frac{\sigma^2\mathbf{v}_k^2}{\sigma_e^2} > 4t_0 + 12t_0^2$. Therefore

$$\begin{aligned} h\left(\frac{\sigma^2\mathbf{v}_k^2}{\sigma_e^2}\right) &\ge h(4t_0 + 12t_0^2) \ge 1 + 4t_0 + 12t_0^2 - 2t_0\sqrt{1 + 8t_0 + 24t_0^2} \\ &\ge 1 + 4t_0 + 12t_0^2 - 2t_0(1 + \sqrt{24}t_0) \\ &= 1 + 2t_0 + 2(6 - \sqrt{24})t_0^2 \ge 1 + 2t_0 + 2t_0^2 \ . \end{aligned}$$

In summary, we have that for any $k \in I_1$,

$$P\left(\frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} \le \sigma_e^2(1 + 2t_0 + 2t_0^2)\right) \le \exp(-t_0^2 n) \ ,$$

and hence, by union bound, $P(\mathcal{G}_2^c) \le \sum_{k \in I_1} \exp(-t_0^2 n) = s_1\delta/p$. Thus $P(\mathcal{G}_1^c) + P(\mathcal{G}_2^c) \le \delta$.

For $\mathcal{G}_3$, we can simply use Lemma II.14 and get that

$$P(\sigma_{\max}(\boldsymbol{\epsilon}_{\cdot I_0}) > \sigma_e(\sqrt{n} + \sqrt{s} + t_1)) \le \exp(-t_1^2/2) \ .$$

Similarly, we can prove that $P(\mathcal{G}_4^c) \le \exp(-t_1^2/2)$.

For $\mathcal{G}_5$, notice that $\frac{1}{\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0} \sim \mathbf{Norm}(\mathbf{0}_s, \sigma_e^2\mathbf{I}_s)$. Thus $\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2^2 \sim \sigma_e^2\chi_s^2$ and, by Corollary II.11,

$$P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 > \sigma_e(\sqrt{s} + t_2)\right) = P(\sqrt{\chi_s^2} > \sqrt{s} + t_2) \le \exp(-t_2^2/2) \ .$$

To bound the probability of $\mathcal{G}_6$, consider a fixed $k \in I_1^c$, and for this k, consider a fixed set $I_1 \subseteq I' \subseteq I_0$ such that $k \notin I'$. There are at most $(p - s_1)$ choices for k and for each k, there are at most $(s - s_1)!$ choices for I' (if $k \in I_0$, then only $(s - s_1 - 1)!$ choices).

Fix this pair of (k, I'), then $\hat{\mathbf{u}}^{(I')}$ is a random unit-norm vector decided by $\boldsymbol{\epsilon}_{\boldsymbol{\cdot} I'}$ and is independent of $\boldsymbol{\epsilon}_{\boldsymbol{\cdot} k}$ (since $k \notin I'$, and $\Sigma_e = \sigma_e^2 \mathbf{I}_p$). If we condition on $\mathbf{X}_{\boldsymbol{\cdot} I'}$, the distribution of $\boldsymbol{\epsilon}_{\boldsymbol{\cdot} k}$ will not change, but $\hat{\mathbf{u}}^{(I')}$ becomes a constant vector and thus $\hat{\mathbf{u}}^{(I')T} \boldsymbol{\epsilon}_{\boldsymbol{\cdot} k} \,\Big|\, \mathbf{X}_{\boldsymbol{\cdot} I'} \sim \mathbf{Norm}(0, \sigma_e^2)$. Since the conditional distribution does not depend on $\mathbf{X}_{\boldsymbol{\cdot} I'}$, we know the marginal distribution is also $\hat{\mathbf{u}}^{(I')T} \boldsymbol{\epsilon}_{\boldsymbol{\cdot} k} \sim \mathbf{Norm}(0, \sigma_e^2)$. Thus

$$\begin{aligned} & P\left(\left|\boldsymbol{\epsilon}_{\boldsymbol{\cdot} k}^T \hat{\mathbf{u}}^{(I')}\right| > \sigma_e \sqrt{2 \log p + 2 s \log s)} + t_3\right) \\ =& P\left(|Z| > \sqrt{2 \log p + 2 s \log s} + t_3\right) \ . \end{aligned}$$

Now we have no more than $p \times s!$ pairs of (k, I'). By Sterling's inequality $\log(p \times s!) \leq \log p + s \log p$. Thus, by Lemma II.13,

$$P(\mathcal{G}_6^c) \leq -2 \exp(-t_3^2/2) \ .$$

□

Note that for some values of $(n, p, s, \sigma_e, \sigma)$; for example, if $\sigma/\sigma_e < 2$, then $\kappa_2 < 0$ no matter how large n is. Otherwise, it is easy to use the result as a sample size calculator. For given (p, s, σ_e, σ) and probability, we can solve for n such that the differentiable signal levels κ_1, κ_2 and the error bound $\|\hat{\mathbf{v}} - \mathbf{v}\|_2$ are as small as needed.

The above theorem is very general and totally non-asymptotic, which makes it hard to understand. Depending on the scale of the problem parameters $(n, p, s, \sigma_e, \sigma)$, one can choose proper values for δ, t_1, t_2, t_3 and transform the non-asymptotic theory into various asymptotic forms. The following corollary is an example.

Corollary II.5. *Denote $\rho = \sigma^2/\sigma_e^2$. Suppose ρ is large enough, and $\frac{s\log p}{\rho n}, \frac{s^2 \log p}{\rho^2 n}, \frac{s^2 \log s}{\rho n} \leq c$ for some small enough c. Then, with probability $1 - C' \exp(-c's) - C'' p^{-c''}$,*

$$\left\|\hat{\mathbf{v}}^{AT} - \mathbf{v}\right\|_2 \leq C \left(\frac{\sqrt{s \log p}}{\sqrt{n\rho}} + \frac{\sqrt{s^2 \log s}}{\sqrt{n\rho}} \right) . \tag{2.27}$$

for some constant C. Moreover, simultaneously the estimated sparse set $\tilde{I}^{AT}$ excludes all coordinates so that $v_k = 0$ while includes all coordinates so that $|v_k| \geq \tilde{C}\sqrt{\frac{\log p + s \log s}{n\rho}}$ for some $\tilde{C}$.

Proof. Let $\delta = 1/p^\alpha$, then $\frac{\log p - \log \delta}{n} = (1+\alpha)\log p/n$. Thus, we have

$$\begin{aligned}
\sqrt{s}\kappa_1 =\leq & \sqrt{\frac{s}{\rho}} \left(2 \left(\frac{\log p - \log \delta}{n} \right)^{1/4} + \sqrt{12} \left(\frac{\log p - \log \delta}{n} \right)^{1/2} \right) \\
& \leq 2(1+\alpha)^{1/4} \left(\frac{s^2 \log p}{\rho^2 n} \right)^{1/4} + \sqrt{12(1+\alpha)} \left(\frac{s \log p}{\rho n} \right)^{1/2} \\
& \leq 2((1+\alpha)c)^{1/4} + \sqrt{12c(1+\alpha)} =: c_1 .
\end{aligned}$$

Let $t_1 = \alpha_1 \sqrt{s}$, $t_2 = \alpha_2 \sqrt{s}$, $t_3 = \alpha_3 \sqrt{\log p + s \log s}$, then

$$\begin{aligned}
\kappa_2 \leq \kappa_2' \leq & \frac{1 + \frac{1}{\sqrt{\rho}} + (1+\alpha_1)\sqrt{\frac{s}{n\rho}}}{\sqrt{1-\epsilon_1^2} - \frac{2}{\sqrt{\rho}} - (1 + 2\alpha_1 + \alpha_2)\sqrt{\frac{s}{n\rho}}} 2(\sqrt{2} + \alpha_3) \sqrt{\frac{\log p + s \log s}{n\rho}} \\
\leq & \frac{1 + \frac{1}{\sqrt{\rho}} + (1+\alpha_1)c}{\sqrt{1-c_1^2} - \frac{2}{\sqrt{\rho}} - (1 + 2\alpha_1 + \alpha_2)c} 2(\sqrt{2} + \alpha_3) \sqrt{\frac{\log p + s \log s}{n\rho}} \\
=: & c_2' \sqrt{\frac{\log p + s \log s}{n\rho}} ,
\end{aligned}$$

and hence

$$\sqrt{s}\kappa_2 \leq c_2' \left(\sqrt{\frac{s \log p}{n\rho}} + \sqrt{\frac{s^2 \log s}{n\rho}} \right) \leq 2c_2' c =: c_2 .$$

Therefore,

$$\begin{aligned}
\left\|\hat{\mathbf{v}}^{(AT)}-\mathbf{v}\right\|_2 &\leq \sqrt{2}s\kappa_2+\frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_2^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right)\\
&\quad+\frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_2^2)}\left[\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2-1\right]\\
&\leq\left(\sqrt{2}c_2'+\frac{4\sqrt{2}(1+\alpha_2)}{\sqrt{1-c_2^2}}\right)\left(\sqrt{\frac{s\log p}{n\rho}}+\sqrt{\frac{s^2\log s}{n\rho}}\right)\\
&\quad+\frac{4\sqrt{2}(1+\alpha_1)}{1-c_2^2}\sqrt{\frac{s}{n}\frac{1}{\rho}}+\frac{2\sqrt{2}(1+\alpha_1)^2}{1-c_2^2}\frac{s}{n\rho}\\
&=: c_3\left(\sqrt{\frac{s\log p}{n\rho}}+\sqrt{\frac{s^2\log s}{n\rho}}\right)+c_4\sqrt{\frac{s}{n}\frac{1}{\rho}}+c_5\frac{s}{n\rho}\ .
\end{aligned}$$

Since we assume ρ to be large enough, $\frac{1}{\rho}=O(1)$; also $\frac{s\log p}{n\rho}=O(1)$. Thus the last two terms are both $O(\sqrt{\frac{s\log p}{n\rho}})$ and hence there exists some constant C such that

$$\left\|\hat{\mathbf{v}}^{(AT)}-\mathbf{v}\right\|_2\leq C\left(\sqrt{\frac{s\log p}{n\rho}}+\sqrt{\frac{s^2\log s}{n\rho}}\right)\ .$$

For small enough c and large enough ρ, we can find $\alpha,\alpha_1,\alpha_2,\alpha_3$, so that $c_1,c_2\in(0,1)$. The probability is

$$\begin{aligned}
&1-p^{-\alpha}-2\exp(-\alpha_1^2 s/2)-\exp(-\alpha_2^2 s/2)-2\exp(-\alpha_3(\log p+s\log s)/2)\\
\geq&1-C'\exp(-c's)-C''p^{-c'}
\end{aligned}$$

for some C',C'',c',c''. With this probability, we achieve $\kappa_2\leq\tilde{C}\sqrt{\frac{\log p+s\log s}{n\rho}}$, and $\left\|\hat{\mathbf{v}}^{AT}-\mathbf{v}\right\|_2\leq C\left(\sqrt{\frac{s\log p}{n\rho}}+\sqrt{\frac{s^2\log s}{n}}\right)$ for some $C,\tilde{C}$. All these constants depend on ρ,c. □

If σ,σ_e,s are fixed while p increase with n, then SPCA-DT can differentiate signals of rate $(\log p/n)^{1/4}$, while SPCA-AT can differentiate signals of rate $(\log p/n)^{1/2}$, which matches the results in the literature. An example of such scenario is when $\mathbf{v}=(\mathbf{v}_s,0,...,0)$, where $\mathbf{v}_s\in\mathbb{R}^s$ is fixed, and $p-s$ coordinates are completely noise. It is still restrictive, and indeed in the PCA literature, the sparsity is often not in ℓ_0

sense so that $\mathbf{v}$ is allowed to have many small non-zero elements. We assume exact sparsity, but the proof we provide here is modular and by modifying some modules, it can be applied to other models.

Remark II.6 (Technical bottleneck)**.** *There is a condition that $s^2 \log s/n\rho$ is bounded, and a $\sqrt{s^2 \log s/n}$ term in the error, which is not ideal. The extra s in the asymptotic rate comes from $\mathcal{G}_6$, which is to bound $\max_{k \notin \hat{I}} \left\| \boldsymbol{\epsilon}_{\cdot k}^T \hat{\mathbf{u}}^{(\hat{I})} \right\|_2 / n$, where $\hat{I}$ is the DT subset and $\hat{\mathbf{u}}^{(\hat{I})}$ is the top left singular vector of $\mathbf{X}_{\cdot \hat{I}}$. This is achieved using union bound of $(p - s_1) \times (s - s_1)!$ Gaussian tails, which is not tight.*

The difficulty is due to the fact that $\hat{I}$ is random, and $\boldsymbol{\epsilon}_{\cdot k}, \hat{\mathbf{u}}^{(\hat{I})}$ are correlated with this random set; if $\hat{I}$ is a constant set, then $\boldsymbol{\epsilon}_{\cdot k}^T \hat{\mathbf{u}}^{(\hat{I})}$ is a standard Gaussian variable, so we only have at most $p - s_1$ Gaussian tails to bound, and the extra term can be avoided.

Another way to avoid the extra term is to use a sample splitting trick: randomly split the sample into two halves, get the $\hat{I}$ from one half, and run the second thresholding on the other half. However, such approach is unnatural and artificial, and practically not favorable since samples are not used efficiently.

In fact, it can be proved that given $\hat{I}$, $\boldsymbol{\epsilon}_{\cdot k}$ is independent of $\hat{\mathbf{u}}^{(\hat{I})}$ for any $k \notin \hat{I}$. The problem is distribution of $\boldsymbol{\epsilon}_{\cdot k}$ after conditioning is no longer normal, and hard to characterize.

2.4.3 Single-spike and non-i.i.d. noises, $D = 1$, $\Sigma_e \neq \sigma_e^2 \mathbf{I}_p$

In this section we still assume the single-spike model

$$\mathbf{X} = \sigma \mathbf{u} \mathbf{v}^T + \boldsymbol{\epsilon} \ , \tag{2.28}$$

but $\boldsymbol{\epsilon}$ no longer has i.i.d. entries; instead we assume that rows of $\boldsymbol{\epsilon}$ i.i.d. follow $\mathbf{Norm}(\mathbf{0}_p, \Sigma_e)$. We are interested in this model because the PCA step in sliced inverse regression (SIR) does not have i.i.d. noise variables.

Similar to the i.i.d. case, we have the following theorem which shows that, with high probability, SPCA-DT and SPCA-AT exclude all coordinates that have zero signals while including all coordinates whose signals are above certain level.

Theorem II.7. *Assume the single-spike model where the noise* $\boldsymbol{\epsilon}$ *has i.i.d. rows that follow* $\mathbf{Norm}(\mathbf{0}_p, \Sigma_e)$, *so that* $\|\Sigma_e\|_{\mathrm{op}} \le \sigma_e^2$. *Suppose that there exist constants* $0 < \delta < 1$ *and* $t_1, t_2, t_3, t_4 > 0$ *such that*

$$t_0 = \sqrt{\frac{\log p - \log \delta}{n}}\ ,$$

$$\kappa_1 = \frac{2\sigma_e}{\sigma}\sqrt{1 + 2t_0 + 2t_0^2}\ ,$$

$$\kappa_2 = \kappa_1 \wedge \kappa_2', \quad \kappa_2' = \frac{2\sigma_e}{\sigma}\frac{\frac{\sqrt{2\log p}+t_3}{\sqrt{n}} + \frac{\sigma_e}{\sigma}\left(1 + \frac{\sqrt{s+1}+\sqrt{2\log p}+t_4}{\sqrt{n}}\right)^2}{\sqrt{1 - s\kappa_1^2 - 2\frac{\sigma_e}{\sigma} - \frac{\sigma_e}{\sigma}\frac{3\sqrt{s}+2t_1+t_2}{\sqrt{n}}}}$$

satisfies $\kappa_2 > 0$, $\sqrt{s}\kappa_1 < 1$. *Denote* $I_1 = \{j : |\mathbf{v}_j| > \kappa_1\}$, $I_2 = \{j : |\mathbf{v}_j| > \kappa_2\}$. *Then using the thresholds*

$$\gamma_1 = \sigma_e^2(1 + 2t_0 + 2t_0^2)\ ,$$

$$\gamma_2 = \sigma_e\sigma\frac{\sqrt{2\log p} + t_3}{\sqrt{n}} + \sigma_e^2\frac{(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)^2}{n}\ ,$$

we have

1. *with probability at least* $1 - \delta$, *the DT estimated set* $\hat{I}$ *satisfies*

$$I_1 \subseteq \hat{I} \subseteq I_0\ ;$$

2. *with probability at least* $1 - \delta - 2\exp(-t_1^2/2) - \exp(-t_2^2/2)$ *the DT estimated eigenvector* $\hat{\mathbf{v}}^{DT}$ *satisfies*

$$\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \le \sqrt{2s}\kappa_1 + \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1 - s\kappa_1^2}}\left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1 - s\kappa_1^2)}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2\ ,$$

where $\hat{\mathbf{v}}^{DT}$ *is the one between* $\pm\hat{\mathbf{v}}^{DT}$ *that is closer to* $\mathbf{v}$;

3. *with probability at least* $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)-\exp(-t_4^2/2)$, *the AT estimated set* $\tilde{I}$ *satisfies*
$$I_2 \subseteq \tilde{I} \subseteq I_0 \ ;$$

4. *with probability at least* $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)-\exp(-t_4^2/2)$, *the AT estimated eigenvector* $\hat{\mathbf{v}}^{AT}$ *satisfies*
$$\left\|\hat{\mathbf{v}}^{AT}-\mathbf{v}\right\|_2 \leq \sqrt{2s}\kappa_2+\frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_2^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right)+\frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_2^2)}\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 ,$$
where $\hat{\mathbf{v}}^{AT}$ *is the one between* $\pm\hat{\mathbf{v}}^{AT}$ *that is closer to* $\mathbf{v}$*.*

Proof. 1. **Preparation:**

Define the following random events

$$\mathcal{G}_1 : \max_{k\in I_0^c} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} < \sigma_e^2(1+2t_0+2t_0^2)$$

$$\mathcal{G}_2 : \min_{k\in I_1} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} > \sigma_e^2(1+2t_0+2t_0^2)$$

$$\mathcal{G}_3 : \sigma_{\max}(\boldsymbol{\epsilon}_{\cdot I_0}) \leq \sigma_e(\sqrt{n}+\sqrt{s}+t_1)$$

$$\mathcal{G}_4 : \sigma_{\min}(\boldsymbol{\epsilon}_{\cdot I_0}) \geq \sigma_e(\sqrt{n}-\sqrt{s}-t_1)$$

$$\mathcal{G}_5 : \left\|\frac{\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0}}{\sqrt{n}}\right\|_2 \leq \sigma_e(\sqrt{s}+t_2)$$

$$\mathcal{G}_6 : \max_{k\in I_1^c}\left|\frac{\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}}\right| \leq \sigma_e(\sqrt{2\log p}+t_3)$$

$$\mathcal{G}_7 : \max_{k\in I_1^c}\left\|\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 \leq \sigma_e^2(\sqrt{n}+\sqrt{s+1}+\sqrt{2\log p}+t_4)^2$$

Use $\hat{\sigma}^{(I)}$, $\hat{\mathbf{u}}^{(I)}$ and $\hat{\mathbf{v}}^{(I)}$ to denote the top singular value, the top left singular vector and the top right singular vector of $\mathbf{X}_{\cdot I}$ for a subset of indices I, respectively.

2. **Diagonal thresholding set:**

   This part is the same as Theorem II.4. If $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, then obviously, by setting $\gamma_1 = \sigma_e^2(1 + 2t_0 + 2t_0^2)$, we have the DT subset $\hat{I}$ satisfies $I_1 \subseteq \hat{I} \subseteq I_0$. Therefore

$$P(I_1 \subseteq \hat{I} \subseteq I_0) \geq 1 - P(\mathcal{G}_1^c) - P(\mathcal{G}_2^c)\ . \tag{2.29}$$

3. **From the DT set to the corresponding estimator:**

   This part is almost the same as Theorem II.4. Assume $\mathcal{G}_1$ to $\mathcal{G}_5$ all hold. We still have

$$\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \leq \left\|\hat{\mathbf{v}}^{(\hat{I})} - \frac{\mathbf{v}_{\hat{I}}}{\left\|\mathbf{v}_{\hat{I}}\right\|_2}\right\|_2 + \left\|\begin{pmatrix}\mathbf{v}_{\hat{I}} / \left\|\mathbf{v}_{\hat{I}}\right\|_2 \\ \mathbf{0}\end{pmatrix} - \mathbf{v}\right\|_2\ . \tag{2.30}$$

   The second term is the bias term that satisfies

$$\left\|\begin{pmatrix}\mathbf{v}_{\hat{I}} / \left\|\mathbf{v}_{\hat{I}}\right\|_2 \\ \mathbf{0}\end{pmatrix} - \mathbf{v}\right\|_2 \leq \sqrt{2}\left\|\mathbf{v}_{I_1}^c\right\| \tag{2.31}$$

   For the variance term, we still use a perturbation bound of top eigenvector with a slightly different breakdown. Note that $\hat{\mathbf{v}}_{\hat{I}}$ is the top eigenvector of

$$(\sigma\mathbf{u}\mathbf{v}_{\hat{I}}^T + \boldsymbol{\epsilon}_{.\hat{I}})^T(\sigma\mathbf{u}\mathbf{v}_{\hat{I}}^T + \boldsymbol{\epsilon}_{.\hat{I}})/n = (\sigma^2\mathbf{v}_{\hat{I}}\mathbf{v}_{\hat{I}}^T) + (M_1 + M_1^T + M_2)$$

   where

$$M_1 = \sigma\mathbf{v}_{\hat{I}}\mathbf{u}^T\boldsymbol{\epsilon}_{.\hat{I}}/n$$

$$M_2 = (\boldsymbol{\epsilon}_{.\hat{I}}^T\boldsymbol{\epsilon}_{.\hat{I}}^T/n)\ .$$

   We can view $M_1 + M_1^T + M_2$ as the perturbation term; $\sigma^2\mathbf{v}_{\hat{I}}\mathbf{v}_{\hat{I}}^T$ is the signal term, whose leading eigenvector is $\frac{\mathbf{v}_{\hat{I}}}{\left\|\mathbf{v}_{\hat{I}}\right\|_2}$ and the top two eigenvalues are $\sigma^2\left\|\mathbf{v}_{\hat{I}}\right\|_2^2$

and 0, so the eigenvalue gap is $\sigma^2 \left\|\mathbf{v}_{\hat{I}}\right\|_2^2 \geq \sigma^2 \left\|\mathbf{v}_{I_1}\right\|_2^2$. Thus by Davis Kahan's inequality (Lemma II.18), we have

$$\left\|\hat{\mathbf{v}}^{(\hat{I})} - \frac{\mathbf{v}_{\hat{I}}}{\left\|\mathbf{v}_{\hat{I}}\right\|_2}\right\|_2 \leq \frac{2\sqrt{2}\left\|M_1 + M_1^T + M_2\right\|_{\mathrm{op}}}{\sigma^2 \left\|\mathbf{v}_{I_1}\right\|_2^2} . \tag{2.32}$$

Since M_1 is of rank 1,

$$\left\|M_1\right\|_{\mathrm{op}} = \sigma \left\|\mathbf{v}_{\hat{I}}\right\|_2 \left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}}^T\right\|_2 / n .$$

Note that $\hat{I} \subseteq I_0$, so $\left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}}^T\right\|_2 \leq \left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot I_0}^T\right\|_2$. By $\mathcal{G}_5$,

$$\left\|M_1\right\|_{\mathrm{op}} \leq \sigma\sigma_e \left\|\mathbf{v}_{\hat{I}}\right\|_2 (\sqrt{s} + t_2)/\sqrt{n} . \tag{2.33}$$

As a sub-block of $\boldsymbol{\epsilon}_{\cdot I_0}^T \boldsymbol{\epsilon}_{\cdot I_0}^T / n$, M_2 has operator norm smaller than that of the whole matrix. Thus, by $\mathcal{G}_3$,

$$\left\|M_2\right\|_{\mathrm{op}} \leq \sigma_e^2 \left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 . \tag{2.34}$$

Substituting (2.33) and (2.34) into (2.32), we get

$$\begin{aligned}\left\|\hat{\mathbf{v}}^{(\hat{I})} - \frac{\mathbf{v}_{\hat{I}}}{\left\|\mathbf{v}_{\hat{I}}\right\|_2}\right\|_2 \leq & \frac{4\sqrt{2}\sigma_e}{\sigma \left\|\mathbf{v}_{I_1}\right\|_2}\left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ & + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2 \left\|\mathbf{v}_{I_1}\right\|_2^2}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 .\end{aligned} \tag{2.35}$$

In summary, substituting (2.31) and (2.35) into (2.30), and using the fact that $\left\|\mathbf{v}_{I_1^c}\right\|_2 \leq \sqrt{s}\kappa_1$, we have that, with probability at least $1 - \sum_{j=1,2,3,4,5} P(\mathcal{G}_j^c)$,

$$\begin{aligned}\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \leq \sqrt{2s}\kappa_1 & + \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1 - s\kappa_1^2}}\left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ & + \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1 - s\kappa_1^2)}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 .\end{aligned} \tag{2.36}$$

4. **Second thresholded set:**

For any $k \leq p$ and $I \subseteq \{1, 2, ..., p\}$, define

$$\Delta_{kI} = \boldsymbol{\epsilon}_{\cdot k}^T \mathbf{X}_{\cdot I} / n ,$$

$$S_{kI} = \mathbf{X}_{\cdot k}^T \mathbf{X}_{\cdot I}/n = \sigma \mathbf{v}_k(\sigma \mathbf{v}_I^T/n + \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot I}/n) + \Delta_{kI} \ .$$

Assume $\mathcal{G}_1$-$\mathcal{G}_7$ hold, then as, in Theorem II.4, we have for $k \in I_0$ and $l \in I_0^c$ that

$$w_l = \left|\Delta_{l\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}\right| \ , \tag{2.37}$$

$$\begin{aligned} w_k &= \left|\sigma \mathbf{v}_k \left(\sigma \left\|\mathbf{v}_{\hat{I}}\right\|_2 \times \left(\frac{\mathbf{v}_{\hat{I}}}{\|\mathbf{v}_{\hat{I}}\|}\right)^T \hat{\mathbf{v}}^{(\hat{I})} + \mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}} \hat{\mathbf{v}}^{(\hat{I})}/n\right) + \Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}\right| \\ &\geq \sigma \mathbf{v}_k \left(\sigma \left\|\mathbf{v}_{\hat{I}}\right\|_2 \left|\left(\frac{\mathbf{v}_{\hat{I}}}{\|\mathbf{v}_{\hat{I}}\|}\right)^T \hat{\mathbf{v}}^{(\hat{I})}\right| - \left\|\mathbf{u}^T \boldsymbol{\epsilon}_{\cdot \hat{I}}/n\right\|_2\right) - \left|\Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}\right| \\ &\geq \sigma \mathbf{v}_k \left(\sigma \sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e \frac{3\sqrt{s} + 2t_1 + t_2}{\sqrt{n}}\right) - \left|\Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}\right| \ . \end{aligned} \tag{2.38}$$

The noise term $|\Delta_{k\hat{I}} \mathbf{v}^{(\hat{I})}|$ will be bounded differently. Note that

$$\Delta_{k\hat{I}} = \frac{\sigma}{n} \boldsymbol{\epsilon}_{\cdot k}^T \mathbf{u} \mathbf{v}_{\hat{I}} + \frac{1}{n} \boldsymbol{\epsilon}_{\cdot k} \boldsymbol{\epsilon}_{\cdot \hat{I}} \ .$$

Thus, since $I_1 \subseteq \hat{I} \subseteq I_0$, we have

$$\begin{aligned} |\Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}| &\leq \left|\frac{\sigma}{n} \boldsymbol{\epsilon}_{\cdot k}^T \mathbf{u}\right| \left|\mathbf{v}_{\hat{I}}^T \hat{\mathbf{v}}_{\hat{I}}\right| + \left\|\frac{1}{n} \boldsymbol{\epsilon}_{\cdot k}^T \boldsymbol{\epsilon}_{\cdot \hat{I}}\right\|_2 \|\hat{\mathbf{v}}_{\hat{I}}\|_2 \\ &\leq \frac{\sigma}{\sqrt{n}} \left|\frac{\boldsymbol{\epsilon}_{\cdot k}^T \mathbf{u}}{\sqrt{n}}\right| + \left\|\frac{1}{n} \boldsymbol{\epsilon}_{\cdot k}^T \boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 \ . \end{aligned}$$

By $\mathcal{G}_6$, the first term is bounded by $\sigma_e \sigma(\sqrt{2 \log p} + t_3)/\sqrt{n}$. By $\mathcal{G}_7$, the second term above is bounded by $\sigma_e^2(\sqrt{n} + \sqrt{s+1} + \sqrt{2 \log p} + t_4)^2/n$. Thus, if we let

$$\gamma_2 = \sigma_e \sigma(\sqrt{2 \log p} + t_3)/\sqrt{n} + \sigma_e^2(\sqrt{n} + \sqrt{s+1} + \sqrt{2 \log p} + t_4)^2/n \ ,$$

then

$$\begin{gathered} w_l < \gamma_2, \quad l \in I_0^c \ , \\ w_k \geq \sigma \mathbf{v}_k \left(\sigma \sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e \frac{3\sqrt{s} + 2t_1 + t_2}{\sqrt{n}}\right) - \gamma_2, \quad k \in I_0 \backslash \hat{I} \ . \end{gathered}$$

If we let

$$\begin{aligned} \kappa_2' &= \frac{2\gamma_2}{\sigma\left(\sigma \sqrt{1 - s\kappa_1^2} - 2\sigma_e - \sigma_e \frac{3\sqrt{s} + 2t_1 + t_2}{\sqrt{n}}\right)} \\ &= \frac{2\sigma_e}{\sigma} \frac{\frac{\sqrt{2 \log p} + t_3}{\sqrt{n}} + \frac{\sigma_e}{\sigma}\left(1 + \frac{\sqrt{s+1} + \sqrt{2 \log p} + t_4}{\sqrt{n}}\right)^2}{\sqrt{1 - s\kappa_1^2} - 2\frac{\sigma_e}{\sigma} - \frac{\sigma_e}{\sigma} \frac{3\sqrt{s} + 2t_1 + t_2}{\sqrt{n}}} \end{aligned}$$

then $\mathbf{v}_k > \kappa_2'$ implies $w_k > \gamma_2$. Thus, with probability at least $1 - \sum_{j\leq 7} P(\mathcal{G}_j^c)$,

$$I_2 \subseteq \tilde{I} \subseteq I_0 \ .$$

5. **From the second set to the corresponding estimator:**

We know that $I_2 \subseteq \tilde{I} \subseteq I_0$, where I_2, I_0 are two fixed sets. We can use the same argument as in part 3, and similar to (2.36), with probability at least $1 - \sum_{j=1}^{7} P(\mathcal{G}_j^c)$,

$$\begin{aligned}\left\|\hat{\mathbf{v}}^{AT} - \mathbf{v}\right\|_2 \leq \sqrt{2s}\kappa_2 &+ \frac{4\sqrt{2}\sigma_e}{\sigma\sqrt{1-s\kappa_2^2}}\left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ &+ \frac{2\sqrt{2}\sigma_e^2}{\sigma^2(1-s\kappa_2^2)}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ .\end{aligned} \tag{2.39}$$

6. **Bound the probabilities:**

The rest of the proof is to bound the probabilities of the random events $\mathcal{G}_1^c, ..., \mathcal{G}_7^c$. For $\mathcal{G}_1$, note that for $k \in I_0^c$, $\mathbf{X}_{\cdot k} = \boldsymbol{\epsilon}_{\cdot k}$. Marginally $\|\boldsymbol{\epsilon}_{\cdot k}\|_2^2 \sim [\Sigma_e]_{kk}\chi_n^2$. Also note that $\|\Sigma_e\|_{\text{op}} \leq \sigma_e^2$ implies $[\Sigma_e]_{kk} \leq \sigma_e^2$. Thus, by Corollary II.11, we have

$$P\left(\frac{\|\boldsymbol{\epsilon}_{\cdot k}\|_2^2}{n} \geq \sigma_e^2(1 + 2t_0 + 2t_0^2)\right) \leq P\left(\chi_n^2/n > 1 + 2t_0 + 2t_0^2\right) \leq \exp(-t_0^2 n) \ ,$$

so by union bound $P(\mathcal{G}_1^c) \leq \sum_{k\in I_0^c} \exp(-t_0^2 n) = (p-s)\delta/p$.

For $\mathcal{G}_2$, note that for $k \in I_1$, we have

$$\|\mathbf{X}_{\cdot k}\|_2 / \sqrt{n} \geq \|\sigma \mathbf{v}_k \mathbf{u}\|_2 / \sqrt{n} - \|\boldsymbol{\epsilon}_{\cdot k}\|_2 / \sqrt{n}$$

and $\|\sigma\mathbf{v}_k\mathbf{u}\|_2/\sqrt{n} = \sigma|\mathbf{v}_k| \geq 2\sigma_e\sqrt{1+2t_0+2t_0^2}$. Thus

$$\begin{aligned}&P\left(\|\mathbf{X}_{\cdot k}\|_2 / \sqrt{n} \leq \sigma_e\sqrt{1 + 2t_0 + 2t_0^2}\right) \\ \leq &P(\|\boldsymbol{\epsilon}_{\cdot k}\|_2 / \sqrt{n} \geq \sigma_e\sqrt{1 + 2t_0 + 2t_0^2}) \leq \exp(-t_0^2 n)\end{aligned}$$

The last inequality is obtained using the same argument as that for $k \in I_0^c$ above. Thus, by union bound, we have $P(\mathcal{G}_2^c) \leq \sum_{k\in I_1} \exp(-t_0^2 n) = s_1\delta/p$, Therefore $P(\mathcal{G}_1^c) + P(\mathcal{G}_2^c) \leq \delta$.

For $\mathcal{G}_3, \mathcal{G}_4$, denote $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_{\cdot I_0}[\Sigma_e]_{I_0 I_0}^{-1/2}$, then $\tilde{\boldsymbol{\epsilon}}$ is a i.i.d. Gaussian ensemble, so by Lemma II.14,

$$P\left(\sigma_{\max}\left(\tilde{\boldsymbol{\epsilon}}\right) > \sqrt{n} + \sqrt{s} + t_1\right) \leq \exp(-t_1^2/2)\ .$$

We know that $\left\|[\Sigma_e]_{I_0 I_0}^{1/2}\right\|_{\text{op}} = \|[\Sigma_e]_{I_0 I_0}\|_{\text{op}}^{1/2} \leq \|\Sigma_e\|_{\text{op}}^{1/2} \leq \sigma_e$, so $\sigma_{\max}\left(\boldsymbol{\epsilon}\right) \leq \sigma_e \sigma_{\max}\left(\tilde{\boldsymbol{\epsilon}}\right)$. Therefore

$$P\left(\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_0}\right) > \sigma_e(\sqrt{n} + \sqrt{s} + t_1)\right) \leq P\left(\sigma_{\max}\left(\tilde{\boldsymbol{\epsilon}}\right) > \sqrt{n} + \sqrt{s} + t_1\right) \leq \exp(-t_1^2/2)\ .$$

We can prove a similar inequality for $\mathcal{G}_4$.

For $\mathcal{G}_5$, notice that $\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}} \sim \mathbf{Norm}(\mathbf{0}_s, \mathbf{I}_s)$, and therefore $\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}}\right\|_2^2 \sim \chi_s^2$. By Corollary II.11,

$$P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}}\right\|_2 > \sqrt{s} + t_2\right) = P(\sqrt{\chi_s^2} > \sqrt{s} + t_2) \leq \exp(-t_2^2/2)\ .$$

Since $\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 \leq \left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}}\right\|_2 \|[\Sigma_e]_{I_0 I_0}\|_{\text{op}} \leq \sigma_e \left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}}\right\|_2$, we have

$$P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 > \sigma_e(\sqrt{s} + t_2)\right) \leq P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{u}^T\tilde{\boldsymbol{\epsilon}}\right\|_2 > \sqrt{s} + t_2\right) \leq \exp(-t_2^2/2)\ .$$

For $\mathcal{G}_6$, notice that $\forall k$, $\frac{\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}} \sim \mathbf{Norm}(0, [\Sigma_e]_{kk})$ and $[\Sigma_e]_{kk} \leq \sigma_e^2$. By Lemma II.13,

$$\begin{aligned} P\left(\max_{k \notin I_1^c}\left|\frac{\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}\sigma_e}\right| > (\sqrt{2\log p - s_1} + t_3)\right) &\leq P\left(\max_{k \notin I_1^c}\left|\frac{\mathbf{u}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n[\Sigma_e]_{kk}}}\right| > \sqrt{2\log p} + t_3\right) \\ &\leq 2\exp(-t_3^2/2)\ . \end{aligned}$$

For $\mathcal{G}_7$, note that $\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot I_0}$ is a sub-block of $\boldsymbol{\epsilon}_{\cdot I_0 \cup \{k\}}^T\boldsymbol{\epsilon}_{\cdot I_0 \cup \{k\}}$, so

$$\left\|\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 \leq \sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_0 \cup \{k\}}\right)^2\ .$$

If $k \in I_0$, then $I_0 \cup \{k\} = I_0$. This has been covered in $\mathcal{G}_3$, and we have

$$\begin{aligned} &P\left(\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_0}\right) > \sigma_e(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)\right) \\ \leq &P\left(\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_0}\right) > \sigma_e(\sqrt{n} + \sqrt{s} + t_4)\right) \leq \exp(-t_4^2/2)\ . \end{aligned}$$

If $k \in I_1$, then the set $I_{0,k} = I_0 \cup \{k\}$ is of size $s+1$. Denote $\tilde{\boldsymbol{\epsilon}}^{(k)} = \boldsymbol{\epsilon}_{\cdot I_{0,k}} \Sigma^{-1/2}_{I_{0,k} I_{0,k}}$. We can prove that $\left\| \Sigma^{1/2}_{I_{0,k} I_{0,k}} \right\|_{\text{op}} \leq \sigma_e$, so

$$\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_{0,k}}\right) \leq \sigma_e \sigma_{\max}\left(\tilde{\boldsymbol{\epsilon}}^{(k)}\right) \ .$$

With the same proof as that for $\mathcal{G}_3$, we can prove that

$$\begin{aligned}
&P\left(\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_{0,k}}\right) \geq \sigma_e(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)\right) \\
\leq& P\left(\sigma_{\max}\left(\tilde{\boldsymbol{\epsilon}}^{(k)}\right) \geq (\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)\right) \\
\leq& \exp(-(\sqrt{2\log p} + t_4)^2/2) \leq \exp(-\log p - t_4^2/2) = \exp(-t_4^2/2)/p \ .
\end{aligned}$$

Use union bound, we get

$$\begin{aligned}
&P\left(\max_{k \in I_1^c} \left\|\boldsymbol{\epsilon}_{\cdot k}^T \boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 > \sigma_e^2(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)^2\right) \\
\leq& \frac{p - s_1}{p} \exp(-t_4^2/2) \leq \exp(-t_4^2)/2 \ .
\end{aligned}$$

$\square$

Similar to the case of i.i.d. noises, we have the following asymptotic theory.

Corollary II.8. *Denote $\rho = \sigma^2/\sigma_e^2$. Suppose ρ is large enough, and no less than 1, and that $\frac{s\log p}{\rho n}, \frac{s\sqrt{s}}{\rho n}, \frac{\sqrt{s}}{\rho} \leq c$ for some small enough c, then with probability $1 - C'\exp(-c's) - C''p^{-c''}$,*

$$\left\|\hat{\mathbf{v}}^{AT} - \mathbf{v}\right\|_2 \leq C\left(\frac{\sqrt{s}}{\rho} + \frac{s\sqrt{s}}{n\rho} + \sqrt{\frac{s\log p}{n\rho}}\right) \tag{2.40}$$

for some constant C, C', C'', c', c''. Moreover, simultaneously the estimated sparse set $\tilde{I}^{AT}$ excludes all coordinates so that $v_k = 0$ while includes all coordinates k such that $|v_k| \geq \tilde{C}\left(\frac{1}{\rho} + \frac{s}{n\rho} + \sqrt{\frac{\log p}{n\rho}}\right)$ for some $\tilde{C}$.

Proof. Let $\delta = p^{-\alpha}$, then $\log\delta = -\alpha\log p$, so

$$\kappa_1 \leq \frac{1}{\sqrt{\rho}}\left(1 + \sqrt{2}\sqrt{\frac{\log p - \log\delta}{n}}\right) = \frac{1}{\sqrt{\rho}} + \sqrt{2(1+\alpha)}\sqrt{\frac{\log p}{n\rho}} \ ,$$

$$\sqrt{s}\kappa_1 \leq \sqrt{\frac{s}{\rho}} + \sqrt{2(1+\alpha)}\sqrt{\frac{s\log p}{n\rho}} \leq (1+\sqrt{2(1+\alpha)})c =: c_1 \ .$$

Let $t_1 = \alpha_1\sqrt{s}$, $t_2 = \alpha_2\sqrt{s}$, $t_3 = \alpha_3\sqrt{\log p}$, $t_4 = \alpha_4\sqrt{\log p}$, then

$$\kappa_2' \leq 2\frac{\frac{2}{n\rho}(n+s+1+(2+\sqrt{\alpha_4})^2\log p) + (\sqrt{2}+\alpha_3)\sqrt{\frac{\log p}{n\rho}}}{\sqrt{1-\epsilon_1^2} - \frac{2}{\sqrt{\rho}} - (3+2\alpha_1+\alpha_2)\sqrt{\frac{s}{n\rho}}} \ .$$

The denominator is larger than $\sqrt{1-c_1^2} - \frac{2}{\sqrt{\rho}} - (3+2\alpha_1+\alpha_2)c$, so it can be bounded away from 0 if ρ is not too small while c is small enough (which further implies that c_1' is small enough). The numerator is of rate $O(\frac{1}{\rho} + \frac{1}{n\rho} + \frac{s}{n\rho} + \frac{\log p}{n\rho} + \sqrt{\frac{\log p}{n\rho}})$. As long as $s \not\to 0$, the term $\frac{1}{n\rho}$ vanishes. Moreover, since we assume $\frac{s\log p}{n\rho} < c$, together with $s \not\to 0$ we have that $\frac{\log p}{n\rho} = O(1)$ so that $\frac{\log p}{n\rho} = O(\sqrt{\frac{\log p}{n\rho}})$. Thus,

$$\kappa_2 \leq \kappa_2' \leq \tilde{C}\left(\frac{1}{\rho} + \frac{s}{n\rho} + \sqrt{\frac{\log p}{n\rho}}\right) \ ,$$

for some constant $\tilde{C}$. Thus,

$$\sqrt{s}\kappa_2 \leq \tilde{C}(2c+\sqrt{c}) =: c_2 \ ,$$

and

$$\begin{aligned}
\left\|\hat{\mathbf{v}}^{AT} - \mathbf{v}\right\|_2 &\leq \sqrt{2}\tilde{C}\left(\frac{\sqrt{s}}{\rho} + \frac{s\sqrt{s}}{n\rho} + \sqrt{\frac{s\log p}{n\rho}}\right) + \frac{2\sqrt{2}}{(1-c_2^2)\rho}\left(1+(1+\alpha_1)\sqrt{\frac{s}{n}}\right)^2 \\
&\quad + \frac{4\sqrt{2}(1+\alpha_2)}{\sqrt{\rho(1-c_2^2)}}\sqrt{\frac{s}{n}} \\
&\leq \sqrt{2}\tilde{C}\left(\frac{\sqrt{s}}{\rho} + \frac{s\sqrt{s}}{n\rho} + \sqrt{\frac{s\log p}{n\rho}}\right) + \frac{4\sqrt{2}}{1-c_2^2}\frac{1}{\rho} + \frac{4\sqrt{2}(1+\alpha_1^2)}{1-c_2^2}\frac{s}{n\rho} \\
&\quad + \frac{4\sqrt{2}(1+\alpha_2)}{\sqrt{(1-c_2^2)}}\sqrt{\frac{s}{n\rho}} \\
&\leq C\left(\frac{\sqrt{s}}{\rho} + \frac{s\sqrt{s}}{n\rho} + \sqrt{\frac{s\log p}{n\rho}}\right) \ .
\end{aligned}$$

for some C. Since c is small enough, we can find $\alpha, \alpha_1, ..., \alpha_4$ so that c_1, c_2 are also small enough and lives in $(0,1)$. Hence the proof can proceed, and the probability of the error bound is at least $1 - C'(-c's) - C''p^{-c''}$, for some C', C'', c', c''. □

Although the i.i.d. case is a special case of the non-i.i.d. case, theoretically speaking, there is a major difference. When comparing the error rates, we can see that in i.i.d. case, the error (2.27) converges to 0 when either sample size n or signal noise ratio ρ goes to infinity, while in non-i.i.d. case, (2.40) converges to 0 if ρ goes to infinity, but does not converge to 0 if sample size n goes to infinity. This is not surprising though, because the population covariance is $\sigma^2 \mathbf{v}\mathbf{v}^T + \Sigma_e$, and the top eigenvector of the population covariance is not always $\mathbf{v}$ for general Σ_e.

Similar to Corollary II.8, we can show that SPCA-DT can differentiate signals of level $\kappa_1 \asymp \frac{1}{\sqrt{\rho}}(1 + \sqrt{\frac{\log p}{n}})$ from noise, and the final error is $\left\|\hat{\mathbf{v}}^{DT} - \mathbf{v}\right\|_2 \asymp \sqrt{\frac{s}{\rho}} + \sqrt{\frac{s \log p}{n\rho}}$. As comparison, SPCA-AT can differentiate signals of level $\kappa_2 \asymp \frac{1}{\rho} + \frac{s}{n\rho} + \sqrt{\frac{\log p}{n\rho}}$, and the error is $\left\|\hat{\mathbf{v}}^{AT} - \mathbf{v}\right\|_2 \asymp \frac{\sqrt{s}}{\rho} + \frac{s\sqrt{s}}{n\rho} + \sqrt{\frac{s \log p}{n\rho}}$. As we have explained, ρ needs to goes to infinity in order for either SPCA-AT or SPCA-DT to be consistent. If $\rho \to \infty$, and moreover if $s = O(n)$, then $\kappa_2 = O(\kappa_1)$. If $\rho \to \infty$ but $n = o(s)$, both $\kappa_2 = O(\kappa_1)$ and $\kappa_1 = O(\kappa_2)$ are possible. The advantage of AT over DT is not as clear as it is in the i.i.d. case.

The extra term $\frac{s\sqrt{s}}{n\rho}$ actually looks unnatural. This term is incurred when bounding $\left|\Delta_{k\hat{I}} \hat{\mathbf{v}}^{(\hat{I})}\right|$ in (2.38). The bound is crude and can possibly be more tight, but we have not achieved that with current proof technique.

2.4.4 Multiple spikes $D > 1$

In this section, we consider the multiple-spike model

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V} + \boldsymbol{\epsilon} \ ,$$

where $\mathbf{U}^T\mathbf{U}/n = \mathbf{V}^T\mathbf{V} = \mathbf{I}_D$, $\mathbf{S} = \mathbf{diag}(\sigma_1, ..., \sigma_D)$ and $\boldsymbol{\epsilon}_{i\cdot} \overset{i.i.d.}{\sim} \mathbf{Norm}(\mathbf{0}_p, \Sigma_e)$. This is an extension of the single-spike model. The proof technique in Theorem II.7 can be used to prove the following theorem.

Theorem II.9. *Consider the multiple-spike model* $\mathbf{X} = \mathbf{USV} + \boldsymbol{\epsilon}$*. Assume the noise* $\boldsymbol{\epsilon}$ *has i.i.d. rows drawn from* $\mathbf{Norm}(\mathbf{0}_p, \Sigma_e)$*, where* $\|\Sigma_e\|_{\text{op}} = \sigma_e^2$*. Denote* $\rho = \sigma_D/\sigma_e$ *and* $\rho_* = \sigma_1/\sigma_D$*. Suppose that exist constants* $0 < \delta < 1$ *and* $t_1, t_2, t_3, t_4 > 0$ *such that* $t_0 = \sqrt{\frac{\log p - \log \delta}{n}}$*, and*

$$\kappa_1 = \frac{2}{\rho}\sqrt{1 + 2t_0 + 2t_0^2}\ ,$$

$$\kappa_2 = \kappa_1 \wedge \kappa_2', \quad \kappa_2' = \frac{2}{\rho}\frac{\rho_* \frac{\sqrt{2\log p} + \sqrt{D} + t_3}{\sqrt{n}} + \frac{1}{\rho}\left(1 + \frac{\sqrt{s+1} + \sqrt{2\log p} + t_4}{\sqrt{n}}\right)^2}{1 - \sqrt{s}\kappa_1 - \frac{2}{\rho} - \frac{1}{\rho}\frac{(2+\rho_*)\sqrt{s} + \rho_*\sqrt{D} + 2t_1 + \rho_* t_2}{\sqrt{n}}}$$

satisfies $\kappa_2 > 0$*,* $\sqrt{s}\kappa_1 < 1$*. Denote* $I_1 = \{j : \|\mathbf{V}_{j\cdot}\|_2 > \kappa_1\}$*,* $I_2 = \{j : \|\mathbf{V}_{j\cdot}\|_2 > \kappa_2\}$*. Then using thresholds*

$$\gamma_1 = \sigma_e^2(1 + 2t_0 + 2t_0^2)\ ,$$

$$\gamma_2 = \frac{\sigma_e\sigma_1(\sqrt{2\log p} + \sqrt{D} + t_3)}{\sqrt{n}} + \sigma_e^2\frac{(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)^2}{n}\ ,$$

we have

1. *with probability at least* $1 - \delta$*, the DT estimated set* $\hat{I}$ *satisfies*

$$I_1 \subseteq \hat{I} \subseteq I_0\ ;$$

2. *with probability at least* $1 - \delta - 2\exp(-t_1^2/2) - \exp(-t_2^2/2)$ *the estimated space satisfies*

$$\mathbf{Dist}_{ave}(\hat{\mathcal{V}}^{DT}, \mathcal{V}) \leq \frac{\sqrt{s}\kappa_1}{\sqrt{D}} + \frac{4\sqrt{2}\sigma_e}{\sigma_D(1 - \sqrt{s}\kappa_1)^2}\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{D}{n}} + \frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma_D^2(1 - \sqrt{s}\kappa_1)^2}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2\ ;$$

3. *with probability at least* $1 - \delta - 2\exp(-t_1^2/2) - \exp(-t_2^2/2) - 2\exp(-t_3^2/2) - \exp(-t_4^2/2)$*, the AT estimated set* $\tilde{I}$ *satisfies*

$$I_2 \subseteq \tilde{I} \subseteq I_0\ ;$$

4. *with probability at least* $1-\delta-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)-\exp(-t_4^2/2)$ *the AT estimated eigenvector* $\hat{\mathbf{v}}^{AT}$ *satisfies*

$$\mathbf{Dist}_{ave}(\hat{\mathcal{V}}^{AT},\mathcal{V}) \le \frac{\sqrt{s}\kappa_2}{\sqrt{D}} + \frac{4\sqrt{2}\sigma_e}{\sigma_D(1-\sqrt{s}\kappa_2)^2}\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{D}{n}}+\frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}\sigma_e^2}{\sigma_D^2(1-\sqrt{s}\kappa_2)^2}\left(1+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 .$$

Proof. The proof is the same as the proof of Theorem II.7. We will emphasize all the differences while omitting the parts that are shared between these two theorems.

1. **Preparation:**

   For easy reference we define all random events that are used, even though only event $\mathcal{G}_2, \mathcal{G}_5, \mathcal{G}_6$ are changed:

$$\mathcal{G}_1 : \max_{k\in I_0^c} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} = \frac{\|\boldsymbol{\epsilon}_{\cdot k}\|_2^2}{n} < \sigma_e^2(1+2t_0+2t_0^2) ,$$

$$\mathcal{G}_2 : \min_{k\in I_1} \frac{\|\mathbf{X}_{\cdot k}\|_2^2}{n} = \frac{\left\|\mathbf{U}\mathbf{S}\mathbf{V}_{k\cdot}^T+\boldsymbol{\epsilon}_{\cdot k}\right\|_2^2}{n} > \sigma_e^2(1+2t_0+2t_0^2) ,$$

$$\mathcal{G}_3 : \sigma_{\max}(\boldsymbol{\epsilon}_{\cdot I_0}) \le \sigma_e(\sqrt{n}+\sqrt{s}+t_1) ,$$

$$\mathcal{G}_4 : \sigma_{\min}(\boldsymbol{\epsilon}_{\cdot I_0}) \ge \sigma_e(\sqrt{n}-\sqrt{s}-t_1) ,$$

$$\mathcal{G}_5 : \left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot I_0}}{\sqrt{n}}\right\|_{\text{op}} \le \sigma_e(\sqrt{s}+\sqrt{D}+t_2) ,$$

$$\mathcal{G}_6 : \max_{k\in I_1^c}\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}}\right\|_{\text{op}} \le \sigma_e(\sqrt{2\log p}+\sqrt{D}+t_3) ,$$

$$\mathcal{G}_7 : \max_{k\in I_1^c}\left\|\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2 \le \sigma_e^2(\sqrt{n}+\sqrt{s+1}+\sqrt{2\log p}+t_4)^2 .$$

   Denote $\hat{\mathbf{V}}^{(I)}$ to be the top D right singular vectors of $\mathbf{X}_{\cdot I}$ for a generic set I and hence, for example, $\hat{\mathbf{V}}^{(DT)} = \begin{pmatrix}\hat{\mathbf{V}}^{(\hat{I})}\\ \mathbf{0}_{p-|\hat{I}|\times D}\end{pmatrix}$ and $\hat{\mathbf{V}}^{(AT)} = \begin{pmatrix}\hat{\mathbf{V}}^{(\tilde{I})}\\ \mathbf{0}_{p-|\tilde{I}|\times D}\end{pmatrix}$.

2. **Diagonal thresholding set:**

   This part is exactly the same in Theorem II.7, so we omit it.

3. **From the DT set to the corresponding estimator:**

Since $\mathbf{Dist}_{ave}$ is constant times Frobenius norm, triangular inequality applies. Thus,

$$\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{DT},\mathbf{V}) \le \mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{(\hat{I})},\mathbf{V}_{\hat{I}\cdot}) + \mathbf{Dist}_{ave}\left(\begin{pmatrix}\mathbf{V}_{\hat{I}\cdot}\\ \mathbf{0}\end{pmatrix},\mathbf{V}\right) . \tag{2.41}$$

Here we use vectors instead of spaces in the arguments for notational convenience. For example, $\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{DT},\mathbf{V}) = \mathbf{Dist}_{ave}(\mathbf{colspan}\langle\hat{\mathbf{V}}^{DT}\rangle,\mathbf{colspan}\langle\mathbf{V}\rangle)$.

For the second term of (2.41), which is the bias term, we have

$$\begin{aligned}\mathbf{Dist}_{ave}\left(\begin{pmatrix}\mathbf{V}_{\hat{I}}\\ \mathbf{0}\end{pmatrix},\mathbf{V}\right)^2 &= 1-\frac{1}{D}\left\|\begin{pmatrix}\mathbf{V}_{\hat{I}}\\ \mathbf{V}_{\hat{I}^c}\end{pmatrix}^T\begin{pmatrix}\mathbf{V}_{\hat{I}}(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{-1/2}\\ \mathbf{0}\end{pmatrix}\right\|_{\mathrm{F}}^2\\ &= 1-\frac{1}{D}\left\|(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{1/2}\right\|_{\mathrm{F}}^2 = 1-\frac{\mathbf{tr}\left(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}}\right)}{D}\\ &= \frac{\|\mathbf{V}_{\hat{I}^c}\|_{\mathrm{F}}^2}{D} \le \frac{\sqrt{s}\kappa_1}{\sqrt{D}} .\end{aligned} \tag{2.42}$$

For the first term of (2.41), which is the variance term, note that $\hat{\mathbf{V}}^{(\hat{I})}$ is the top eigenvectors of

$$(\mathbf{U}\mathbf{S}\mathbf{V}_{\hat{I}\cdot}^T+\boldsymbol{\epsilon}_{\cdot\hat{I}})^T(\mathbf{U}\mathbf{S}\mathbf{V}_{\hat{I}\cdot}^T+\boldsymbol{\epsilon}_{\cdot\hat{I}}) = \mathbf{V}_{\hat{I}\cdot}\boldsymbol{\Lambda}\mathbf{V}_{\hat{I}\cdot}^T + M_1 + M_1^T + M_2 ,$$

where

$$M_1 = \mathbf{V}_{\hat{I}\cdot}\mathbf{S}\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot\hat{I}}/n ,$$

$$M_2 = \boldsymbol{\epsilon}_{\cdot\hat{I}}^T\boldsymbol{\epsilon}_{\cdot\hat{I}}^T/n .$$

The signal matrix $\mathbf{V}_{\hat{I}\cdot}\boldsymbol{\Lambda}\mathbf{V}_{\hat{I}\cdot}^T$ is of rank D and $M_1+M_1^T+M_2$ is the perturbation term. Although $\mathbf{V}_{\hat{I}\cdot}$ is not an orthonormal basis, it spans the same space as the eigenvectors of $\mathbf{V}_{\hat{I}\cdot}\boldsymbol{\Lambda}\mathbf{V}_{\hat{I}\cdot}^T$. By Davis Kahan's inequality (Lemma II.19),

$$\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{(\hat{I})},\mathbf{V}_{\hat{I}\cdot}) = \frac{1}{\sqrt{D}}\left\|\sin\Theta(\hat{\mathbf{V}}^{(\hat{I})},\mathbf{V}_{\hat{I}})\right\|_{\mathrm{F}} \le \frac{2\left\|M_1+M_1^T+M_2\right\|_{\mathrm{op}}}{\sigma_{\min}\left(\mathbf{V}_{\hat{I}\cdot}\mathbf{S}\right)^2}. \tag{2.43}$$

For M_1,

$$\|M_1\|_{\text{op}} \le \|\mathbf{V}_{\hat{I}}\|_{\text{op}} \|\mathbf{S}\|_{\text{op}} \left\|\mathbf{U}^T \boldsymbol{\epsilon}^T_{\cdot \hat{I}}\right\|_{\text{op}} / n \ .$$

Note that $\hat{I} \subseteq I_0$, so $\left\|\mathbf{U}^T \boldsymbol{\epsilon}^T_{\cdot \hat{I}}\right\|_{\text{op}} \le \left\|\mathbf{U}^T \boldsymbol{\epsilon}^T_{\cdot I_0}\right\|_{\text{op}}$. Also $\|\mathbf{V}_{\hat{I}}\|_{\text{op}} \le 1$, $\|\mathbf{S}\|_{\text{op}} \le \sigma_1$. By $\mathcal{G}_5$

$$\|M_1\|_{\text{op}} \le \sigma_1 \sigma_e (\sqrt{s} + \sqrt{D} + t_2)/\sqrt{n} \ . \tag{2.44}$$

As a sub-block of $\boldsymbol{\epsilon}^T_{\cdot I_0} \boldsymbol{\epsilon}^T_{\cdot I_0}/n$, and M_2 has operator norm smaller than that of the whole matrix. Thus, by $\mathcal{G}_3$,

$$\|M_2\|_{\text{op}} \le \sigma_e^2 \left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ . \tag{2.45}$$

To bound $\sigma_{\min}(\mathbf{V}_{\hat{I}\cdot}\mathbf{S})$, first note that by Weyl's inequality and the fact that $\hat{I} \subseteq I_1$, we have

$$\begin{aligned}\sigma_{\min}(\mathbf{V}_{\hat{I}\cdot}) \ge \sigma_{\min}(\mathbf{V}_{I_1\cdot}) &= \sigma_{\min}\left(\mathbf{V} - \begin{pmatrix} \mathbf{0} \\ \mathbf{V}_{I_1^c\cdot} \end{pmatrix}\right) \\ &\ge 1 - \sigma_{\max}(\mathbf{V}_{I_1^c\cdot}) \ge 1 - \left\|\mathbf{V}_{I_1^c\cdot}\right\|_{\text{F}} \ge 1 - \sqrt{s}\kappa_1 \ ,\end{aligned}$$

so

$$\sigma_{\min}(\mathbf{V}_{\hat{I}\cdot}\mathbf{S}) \ge \sigma_D \sigma_{\min}(\mathbf{V}_{\hat{I}}) \ge \sigma_D(1 - \sqrt{s}\kappa_1) \ . \tag{2.46}$$

Substituting (2.44), (2.45) and (2.46) into (2.43), we get

$$\begin{aligned}\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{(\hat{I})}, \mathbf{V}_{\hat{I}\cdot}) \le & \frac{4\sqrt{2}\sigma_e}{\sigma_D(1-\sqrt{s}\kappa_1)^2}\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{D}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ & + \frac{2\sqrt{2}\sigma_e^2}{\sigma_D^2(1-\sqrt{s}\kappa_1)^2}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ .\end{aligned} \tag{2.47}$$

In summary, substituting 2.42 and 2.47 into 2.41, we have, with probability at least $1 - \sum_{j=1,2,3,4,5} P(\mathcal{G}_j^c)$, that

$$\begin{aligned}\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{DT}, \mathbf{V}) \le \frac{\sqrt{s}\kappa_1}{\sqrt{D}} & + \frac{4\sqrt{2}\sigma_e}{\sigma_D(1-\sqrt{s}\kappa_1)^2}\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{D}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ & + \frac{2\sqrt{2}\sigma_e^2}{\sigma_D^2(1-\sqrt{s}\kappa_1)^2}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ .\end{aligned} \tag{2.48}$$

4. **Second thresholded set:**

Assume $\mathcal{G}_1$-$\mathcal{G}_7$ hold. For any $k \le p$ and $I \subseteq \{1, 2, ..., p\}$, denote

$$\Delta_{kI} = \boldsymbol{\epsilon}_{\cdot k}^T \mathbf{X}_{\cdot I}/n \ ,$$

$$S_{kI} = \mathbf{X}_{\cdot k}^T \mathbf{X}_{\cdot I}/n = \mathbf{V}_{k\cdot}\mathbf{S}^2\mathbf{V}_I^T/n + \mathbf{V}_{k\cdot}\mathbf{S}\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot I}/n + \Delta_{kI} \ .$$

Then for $k \in I_0 \backslash \hat{I}$ and $l \in I_0^c$, we have that

$$w_l = \left\| \Delta_{l\hat{I}} \hat{\mathbf{V}}^{(\hat{I})} \right\|_2 \ , \tag{2.49}$$

$$w_k \ge = \left\| \mathbf{V}_{k\cdot}\mathbf{S}^2\mathbf{V}_{\hat{I}}^T \hat{\mathbf{V}}^{(\hat{I})} \right\|_2 - \left\| \mathbf{V}_{k\cdot}\mathbf{S}\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot \hat{I}}}{n} \hat{\mathbf{V}}^{(\hat{I})} \right\| - \left\| \Delta_{l\hat{I}} \hat{\mathbf{V}}^{(\hat{I})} \right\|_2 \ . \tag{2.50}$$

For the first term of (2.50), be aware that $\mathbf{V}_{\hat{I}\cdot}$ is not an orthonormal basis, and needs to be transformed. By Wedin's inequality (Lemma II.17), we have

$$\sigma_{\min}\left(\left(\mathbf{V}_{\hat{I}}(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{-1/2} \right)^T \hat{\mathbf{V}}^{(\hat{I})} \right) \ge 1 - \frac{2\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot \hat{I}}\right)}{\sigma_{\min}\left(\mathbf{U}\mathbf{S}\mathbf{V}_{\hat{I}\cdot}^T\right)} = 1 - \frac{2\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot \hat{I}}\right)}{\sqrt{n}\sigma_{\min}\left(\mathbf{V}_{\hat{I}\cdot}\mathbf{S}\right)} \ .$$

The last equality holds because $\mathbf{U}/\sqrt{n}$ is orthonormal. Since $\mathcal{G}_3$ holds, $\sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot \hat{I}}\right) \le \sigma_{\max}\left(\boldsymbol{\epsilon}_{\cdot I_0}\right) \le \sigma_e(\sqrt{n} + \sqrt{s} + t_1)$. The denominator is bounded in (2.46) as $\sigma_{\min}\left(\mathbf{V}_{\hat{I}\cdot}\mathbf{S}\right) \ge \sigma_D(1 - \sqrt{s}\kappa_1)$. Also note that $\sigma_{\min}\left(\mathbf{S}(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{1/2}\right) = \sigma_{\min}\left(\mathbf{V}_{\hat{I}\cdot}\mathbf{S}\right)$. Therefore,

$$\begin{aligned} & \frac{1}{n}\left\| \mathbf{V}_{k\cdot}\mathbf{S}^2\mathbf{V}_{\hat{I}}^T \hat{\mathbf{V}}^{(\hat{I})} \right\|_2 \\ \ge & \left\| \mathbf{V}_{k\cdot} \right\|_2 \sigma_{\min}\left(\mathbf{S}\right) \sigma_{\min}\left(\mathbf{S}(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{1/2}\right) \sigma_{\min}\left(\left(\mathbf{V}_{\hat{I}}(\mathbf{V}_{\hat{I}}^T\mathbf{V}_{\hat{I}})^{-1/2} \right)^T \hat{\mathbf{V}}^{(\hat{I})} \right) \\ \ge & \sigma_D \left\| \mathbf{V}_{k\cdot} \right\|_2 \sigma_D(1 - \sqrt{s}\kappa_1) \left(1 - \frac{2(\sqrt{n} + \sqrt{s} + t_1)}{\sqrt{n}\sigma_D(1 - \sqrt{s}\kappa_1)} \right) \\ = & \sigma_D \left\| \mathbf{V}_{k\cdot} \right\|_2 \left(\sigma_D(1 - \sqrt{s}\kappa_1) - 2\sigma_e - 2\sigma_e\frac{\sqrt{s} + t_1}{\sqrt{n}} \right) \ . \end{aligned} \tag{2.51}$$

For the second term of (2.50), simply by $\mathcal{G}_5$, we have

$$\begin{aligned}\left\|\mathbf{V}_{k\cdot}\mathbf{S}\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot\hat{I}}}{n}\hat{\mathbf{V}}^{(\hat{I})}\right\|_2 &\le \|\mathbf{V}_{k\cdot}\|_2\,\|\mathbf{S}\|_{\mathrm{op}}\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot\hat{I}}}{n}\right\|_{\mathrm{op}}\left\|\hat{\mathbf{V}}^{(\hat{I})}\right\|_{\mathrm{op}} \\ &\le \sigma_1\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot I_0}}{n}\right\|_{\mathrm{op}}\|\mathbf{V}_{k\cdot}\|_2 \\ &\le \sigma_1\sigma_e(\sqrt{s}+\sqrt{D}+t_2)\,\|\mathbf{V}_{k\cdot}\|_2\ .\end{aligned} \tag{2.52}$$

The third term of (2.50) and w_l, $l\in I_0^c$ can be treated together. We have

$$\Delta_{k\hat{I}} = \frac{1}{n}\boldsymbol{\epsilon}_{\cdot k}^T\mathbf{U}\mathbf{S}\mathbf{V}_{\hat{I}\cdot} + \frac{1}{n}\boldsymbol{\epsilon}_{\cdot k}\boldsymbol{\epsilon}_{\cdot\hat{I}}\ .$$

With $\hat{I}\subseteq I_0$,

$$\begin{aligned}\left\|\Delta_{k\hat{I}}\hat{\mathbf{V}}^{(\hat{I})}\right\|_2 &\le \frac{1}{n}\left\|\boldsymbol{\epsilon}_{\cdot k}^T\mathbf{U}\right\|_2\left\|\mathbf{S}\mathbf{V}_{\hat{I}\cdot}\hat{\mathbf{V}}^{(\hat{I})}\right\|_{\mathrm{op}} + \left\|\frac{1}{n}\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot\hat{I}}\right\|_2\left\|\hat{\mathbf{V}}_{\hat{I}}\right\|_{\mathrm{op}} \\ &\le \frac{\sigma_1}{\sqrt{n}}\left\|\frac{\boldsymbol{\epsilon}_{\cdot k}^T\mathbf{U}}{\sqrt{n}}\right\|_2 + \left\|\frac{1}{n}\boldsymbol{\epsilon}_{\cdot k}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_2\ .\end{aligned}$$

By $\mathcal{G}_6$ and $\mathcal{G}_7$, we then have

$$\begin{aligned}\left\|\Delta_{k\hat{I}}\hat{\mathbf{V}}^{(\hat{I})}\right\|_2 &\le \sigma_e\sigma_1\frac{\sqrt{2\log p}+\sqrt{D}+t_3}{\sqrt{n}} \\ &\quad + \frac{\sigma_e^2(\sqrt{n}+\sqrt{s+1}+\sqrt{2\log p}+t_4)^2}{n}\ .\end{aligned} \tag{2.53}$$

Substituting (2.51), (2.52) and (2.53) into (2.49) and (2.50) and defining

$$\gamma_2 = \frac{\sigma_e\sigma_1(\sqrt{2\log p}+\sqrt{D}+t_3)}{\sqrt{n}} + \sigma_e^2\frac{(\sqrt{n}+\sqrt{s+1}+\sqrt{2\log p}+t_4)^2}{n}\ ,$$

$$\rho = \sigma_D/\sigma_e\ ,$$

$$\rho_* = \sigma_1/\sigma_D\ ,$$

$$\begin{aligned}\kappa_2' &= \frac{2\gamma_2}{\sigma_D\left(\sigma_D(1-\sqrt{s}\kappa_1) - 2\sigma_e - \sigma_e\frac{(2+\rho_*)\sqrt{s}+\rho_*\sqrt{D}+2t_1+\rho_* t_2}{\sqrt{n}}\right)} \\ &= \frac{2}{\rho}\frac{\rho_*\frac{\sqrt{2\log p}+\sqrt{D}+t_3}{\sqrt{n}} + \frac{1}{\rho}\left(1+\frac{\sqrt{s+1}+\sqrt{2\log p}+t_4}{\sqrt{n}}\right)^2}{1-\sqrt{s}\kappa_1-\frac{2}{\rho}-\frac{1}{\rho}\frac{(2+\rho_*)\sqrt{s}+\rho_*\sqrt{D}+2t_1+\rho_* t_2}{\sqrt{n}}}\ ,\end{aligned}$$

then

$$w_l < \gamma_2, \quad l\in I_0^c\ ,$$

$$w_k \geq \gamma_2, \quad k : \mathbf{v}_k > \kappa_2' \ .$$

Thus, with probability at least $1 - \sum_{j\leq 7} P(\mathcal{G}_j^c)$,

$$I_2 \subseteq \tilde{I} \subseteq I_0 \ .$$

5. **From the second set to the corresponding estimator:**

   We know that $I_2 \subseteq \tilde{I} \subseteq I_0$, where I_2, I_0 are two fixed sets. Using the same argument as in part 3 and similar to (2.48), we have, with probability at least $1 - \sum_{j=1}^{7} P(\mathcal{G}_j^c)$, that

$$\begin{aligned}\mathbf{Dist}_{ave}(\hat{\mathbf{V}}^{AT}, \mathbf{V}) \leq & \frac{\sqrt{s}\kappa_2}{\sqrt{D}} + \frac{4\sqrt{2}\sigma_e}{\sigma_D(1-\sqrt{s}\kappa_2)^2}\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{D}{n}} + \frac{t_2}{\sqrt{n}}\right) \\ & + \frac{2\sqrt{2}\sigma_e^2}{\sigma_D^2(1-\sqrt{s}\kappa_2)^2}\left(1 + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ .\end{aligned} \tag{2.54}$$

6. **Bound the probabilities:**

   The events $\mathcal{G}_1, \mathcal{G}_3, \mathcal{G}_4, \mathcal{G}_7$ are exactly the same as in Theorem II.7. We only address the other three events.

   For $\mathcal{G}_2$, when $k \in I_1$, we have

$$\|\mathbf{X}_{\cdot k}\|_2 / \sqrt{n} \geq \left\|\mathbf{U}\mathbf{S}\mathbf{V}_{k\cdot}^T\right\|_2 / \sqrt{n} - \|\boldsymbol{\epsilon}_{\cdot k}\|_2 / \sqrt{n} \ ,$$

   and $\|\mathbf{U}\mathbf{S}\mathbf{V}_{k\cdot}\|_2 / \sqrt{n} \geq \sigma_D \|\mathbf{V}_{k\cdot}\|_2 \geq 2\sigma_e\sqrt{1+2t_0+2t_0^2}$. Thus

$$\begin{aligned} & P\left(\|\mathbf{X}_{\cdot k}\|_2 / \sqrt{n} \leq \sigma_e\sqrt{1+2t_0+2t_0^2}\right) \\ \leq & P\left(\|\boldsymbol{\epsilon}_{\cdot k}\|_2 / \sqrt{n} \geq \sigma_e\sqrt{1+2t_0+2t_0^2}\right) \leq \exp(-t_0^2 n) \ .\end{aligned}$$

   By union bound, $P(\mathcal{G}_2^c) \leq \sum_{k\in I_1} \exp(-t_0^2 n) = s_1\delta/p$.

   For $\mathcal{G}_5$, denote $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_{\cdot I_0}[\Sigma_e]_{I_0 I_0}^{-1/2}$, then $\tilde{\boldsymbol{\epsilon}}$ is i.i.d. standard Gaussian ensemble. Consequently $\frac{1}{\sqrt{n}}\mathbf{U}^T\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^{D\times s}$ also has i.i.d. standard Gaussian entries. Thus by Lemma II.14

$$P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\tilde{\boldsymbol{\epsilon}}\right\|_{\text{op}} > \sqrt{s}+\sqrt{D}+t_2\right) \leq \exp(-t_2^2/2)\ .$$

Since $\left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_{\text{op}} \leq \left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\tilde{\boldsymbol{\epsilon}}\right\|_{\text{op}} \|[\Sigma_e]_{I_0 I_0}\|_{\text{op}} \leq \sigma_e \left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\tilde{\boldsymbol{\epsilon}}\right\|_{\text{op}}$, we have

$$\begin{aligned}&P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot I_0}\right\|_{\text{op}} > \sigma_e(\sqrt{s}+\sqrt{D}+t_2)\right)\\ \leq &P\left(\left\|\frac{1}{\sqrt{n}}\mathbf{U}^T\tilde{\boldsymbol{\epsilon}}\right\|_{\text{op}} > \sqrt{s}+\sqrt{D}+t_2\right) \leq \exp(-t_2^2/2)\ .\end{aligned}$$

For $\mathcal{G}_6$, note that $\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}} \sim \mathbf{Norm}(0, [\Sigma_e]_{kk}\mathbf{I}_D)$ for all k; also note that $[\Sigma_e]_{kk} \leq \sigma_e^2$. By Corollary II.11,

$$\begin{aligned}&P\left(\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}\sigma_e}\right\|_2 > \sqrt{D}+\sqrt{2\log p}+t_3\right)\\ \leq &P\left(\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n[\Sigma_e]_{kk}}}\right\|_2 > \sqrt{D}+\sqrt{2\log p}+t_3\right)\\ \leq &\exp(-(\sqrt{2\log p}+t_3)^2/2) \leq \exp(-t_3^2/2)/p\ .\end{aligned}$$

By union bound,

$$P\left(\max_{k\notin I_1}\left\|\frac{\mathbf{U}^T\boldsymbol{\epsilon}_{\cdot k}}{\sqrt{n}\sigma_e}\right\|_2 > \sqrt{D}+\sqrt{2\log p}+t_3\right) \leq \frac{p-s_1}{p}\exp(-t_3^2/2) \leq \exp(-t_3^2/2)\ .$$

□

The error bounds in Theorem II.9 and the error bounds in Theorem II.7 have the same asymptotic rate as long as the rank D and the condition number $\rho_* = \sigma_1/\sigma_D$ of the low rank signal matrix are both of constant rate. The procedure of transforming the theorem above to an asymptotic version will be almost the same that for $D = 1$, so we omit it here.

2.5 Simulation

In this section we provide some numerical results that compare various GSPCA methods. Data are generated from the spiked-covariance model

$$\mathbf{X} = \rho \sum_{d=1}^{D} \mathbf{u}_d \mathbf{v}_d^T + \boldsymbol{\epsilon} \ ,$$

where $\mathbf{u}_d \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times p}$ have i.i.d. standard Gaussian entries. $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_D) \in \mathbb{R}^{p \times D}$ are generated from the following three models, with model configuration parameters (n, p, ρ).

1. V **Model 1:** Single spike, equal signal levels. $D = 1$, $s = 5$, $v_S = \frac{1}{\sqrt{s}}(1, 1, .., 1)^T$, $\mathbf{V} = \mathbf{V}^{(1)} := \begin{pmatrix} v_S \\ \mathbf{0}_{p-s \times 1} \end{pmatrix}$.
2. V **Model 2:** Single spike, unequal signal levels. $D = 1$, $s = 5$, $v_S = (0.8, 0.8^2, ...0.8^s)$, $v_S = v_S / \left\| v_S \right\|_2$. $\mathbf{V} = \mathbf{V}^{(2)} := \begin{pmatrix} v_S \\ \mathbf{0}_{(p-s) \times 1} \end{pmatrix}$.
3. V **Model 3:** Two spikes, unequal signal levels. $D = 2$, $s = 5$, $v_S = (0.8, 0.8^2, ...0.8^s)$, $v_S = v_S / \left\| v_S \right\|_2$. $\mathbf{V} = \mathbf{V}^{(3)} := \begin{pmatrix} \begin{pmatrix} v_S & \mathbf{0}_s \\ \mathbf{0}_s & v_S \end{pmatrix} \\ \mathbf{0}_{(p-2s) \times 2} \end{pmatrix}$.

All the methods we tested have penalty parameters. In order to avoid over-tuning, for each method we only change one tuning parameter while keep others (if any) fixed. We picked 20 values for the one chosen penalty parameter, so that the range is usually wide enough to include the optimal tuning parameter. We check that by calculating the mean errors from 200 independent runs, and see if the minimal error is not from the largest or smallest penalty parameter values. That's usually the case in our experiments, which is specified below.

For SPCA-DT, we vary $\gamma_1 = 1.2^{(-10,-9,...,9,10)}\left(1 + 2\sqrt{\log(p/n)} + 2\log(p/n)\right)$. For SPCA-AT, we fix γ_1 and vary $\gamma_2 = 1.4^{(-15,-14,...,5)}\,(\sigma^2/n)$; the γ_1 is chosen to be either an "optimal" value (which we get by checking the average results from SPCA-DT), and then 2 times that value. For GSPCA-Reg, we vary $\lambda = 1.2^{(-10,-9,...,10)}/\sqrt{n}$. For GSPCA-Power we vary $\tau = (0.1, ..., 2)\,\|\mathbf{V}\|_{1,2}$. For GSPCA-SDP we fix $\eta = 2$, and vary $\lambda = (2, 4, ..., 40)\sqrt{\log p/n}$. Finally as a bench mark, we use the built-in SPCA method from the `sklearn` package in python, which does not get group sparsity; we vary the tunning parameter $\alpha = (0.1, ..., 2)\sqrt{\log p}$.

In the first batch of experiments, we fix $p = 300$ and vary (n, ρ). We tested the three right vectors $\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \mathbf{V}^{(3)}$ and $(n, \rho) = (40, 5), (160, 5), (10, 10), (40, 10)$, so in total we have 12 configurations. For each configuration, we apply GSPCA-Reg, GSPCA-Power, GSPCA-SDP, SPCA-DT, SPCA-AT(with the optimal and a larger γ_1), SPCA-python, each method having 20 tuning; we run 200 independent experiments and calculate the average error $\mathbf{Dist}_{ave}(\hat{\mathbf{V}}, \mathbf{V})$, so each method has 20 average errors according to the 20 tuning parameter values, and we report the smallest average error. This is to mimic CV, and will tell us the potential performance of the compared methods under proper tuning.

Table 2.1 shows the minimal mean errors of all methods and configurations. Generally speaking for $\mathbf{V}^{(1)}$, where non-zero coordinates have uniform signal level, GSPCA-SDP performs the best, and SPCA-AT has no advantage over SPCA-DT; for $\mathbf{V}^{(2)}, \mathbf{V}^{(3)}$ where the signal levels are uneven, SPCA-AT performs the best. From $D = 1$ to $D = 2$ (i.e. $\mathbf{V}^{(2)}$ to $\mathbf{V}^{(3)}$), GSPCA-SDP deteriorates more than thresholding-based methods; the sklearn built-in method does not seem to be affected, because the sparsity of $\mathbf{V}^{(3)}$ is not actually by groups, so penalizing for group sparsity has no advantage. As for computation, GSPCA-Power, GSPCA-SDP and

V Model	(n, ρ)	Reg	Python	Power	SDP	DT	AT	AT
							(optimial	(large
							γ_1)	γ_1)
$\mathbf{V}^{(1)}$	(40,5)	0.217	0.081	0.079	0.034	0.062	0.062	0.070
	(160,5)	0.059	0.035	0.034	0.017	0.030	0.030	0.030
	(10,10)	0.401	0.085	0.083	0.052	0.065	0.063	0.076
	(40,10)	0.101	0.035	0.034	0.018	0.030	0.030	0.030
$\mathbf{V}^{(2)}$	(40,5)	0.282	0.105	0.103	0.093	0.083	0.067	0.066
	(160,5)	0.080	0.045	0.044	0.051	0.030	0.030	0.030
	(10,10)	0.466	0.109	0.110	0.090	0.073	0.066	0.092
	(40,10)	0.140	0.047	0.046	0.035	0.030	0.030	0.030
$\mathbf{V}^{(3)}$	(40,5)	0.322	0.105	0.127	0.134	0.110	0.098	0.103
	(160,5)	0.096	0.047	0.060	0.062	0.045	0.044	0.044
	(10,10)	0.499	0.110	0.136	0.153	0.113	0.105	0.141
	(40,10)	0.177	0.047	0.060	0.054	0.045	0.045	0.045

Table 2.1: Compare the performance of various SPCA algorithm under a wide range of settings.

GSPCA-DT/AT are all very efficient, while GSPCA-Reg and the sklearn method are much less efficient. Despite that the tuning parameters are chosen heuristically, we can still conclude that thresholding based SPCA methods achieve a good balance between computation and accuracy and work for various models well.

It can also be observed that the larger signal noise ratio ρ is, the smaller sample size n is needed to achieve the same error, and the error roughly depends on $n\rho^2$. This is consistent to our theory. To see this more clearly, we ran a second batch of experiments where we set $n = 10, 40, 160, 640, 2560$ and $\rho = \sqrt{4000/n}$, so that $n\rho^2$ is held constant. We can see a clear trade-off between ρ and n. If ρ is too small, even n is very large the estimator is not consistent (see row ten in Table 2.2); on the other hand, there is a limit that how small n can be for the theory to hold, so that if n is simply too small, the error will also start to increase.

Finally, we run a third batch of experiments where we increase p and see how different methods perform for high dimensional case. We compare GSPCA-Power, GSPCA-SDP, SPCA-DT and SPCA-AT, and set $(n, p) = (35, 150), (40, 300), (45, 600)$, so that $\log p/n$ is almost constant (around 0.14). We test an additional model:

V Model	n	Reg	Python	Power	SDP	DT	AT (optimial γ_1)	AT (large γ_1)
$\mathbf{V}^{(1)}$	10	0.184	0.037	0.036	0.024	0.032	0.032	0.032
	40	0.101	0.035	0.034	0.018	0.030	0.030	0.030
	160	0.059	0.035	0.034	0.017	0.030	0.030	0.030
	640	0.043	0.036	0.035	0.017	0.031	0.031	0.031
	2560	0.046	0.043	0.043	0.016	0.037	0.037	0.037
$\mathbf{V}^{(2)}$	10	0.229	0.049	0.047	0.037	0.032	0.032	0.033
	40	0.140	0.047	0.046	0.035	0.030	0.030	0.030
	160	0.080	0.045	0.044	0.051	0.030	0.030	0.030
	640	0.059	0.049	0.047	0.091	0.032	0.032	0.031
	2560	0.063	0.059	0.058	0.310	0.143	0.039	0.038

Table 2.2: Trade off between sample size and signal. $n\rho^2 = 4000$.

V Model	(n,p)	Power	SDP	DT	AT (optimial γ_1)	AT (large γ_1)
$\mathbf{V}^{(1)}$	(35,150)	0.085	0.042	0.066	0.066	0.081
	(40,300)	0.079	0.034	0.062	0.062	0.069
	(45,600)	0.076	0.029	0.059	0.059	0.067
$\mathbf{V}^{(2)}$	(35,150)	0.103	0.098	0.084	0.072	0.076
	(40,300)	0.103	0.093	0.083	0.067	0.063
	(45,600)	0.103	0.083	0.080	0.066	0.061
$\mathbf{V}^{(3)}$	(35,150)	0.131	0.132	0.115	0.106	0.111
	(40,300)	0.127	0.134	0.110	0.098	0.101
	(45,600)	0.128	0.112	0.107	0.096	0.098
$\mathbf{V}^{(4)}$	(35,150)	0.160	0.162	0.142	0.135	0.140
	(40,300)	0.152	0.151	0.133	0.126	0.128
	(45,600)	0.145	0.139	0.126	0.118	0.117

Table 2.3: Performance when p increases

- V **Model 4:** Multiple spikes, group sparsity. $D = 3$, $v_S = (0.8, 0.8^2, ...0.8^5)$, $v_S = v_S / \|v_S\|_2$. $\mathbf{V} \in \mathbb{R}^{p \times D}$ such that for $j = 1, 2, ..., 5$, $\mathbf{V}_{I_j}. = v_j O_j$, where $I_j = \{(j-1)D + 1, ..., jD\}$, and $O_j \in \mathbb{R}^{D \times D}$ is orthonormal.

The results are shown in Table 2.3. All these methods perform well as p increases, and we can still see that thresholded SPCA methods perform better in the case well non-zero covariates have uneven signal levels. Also AT is quite insensitive to the choice of DT threshold (the errors from fixing γ_1 to be the "optimal" value and a large value are similar).

2.6 Auxiliary lemmas

Here we provide the auxiliary lemmas that are used in this chapter.

2.6.1 Exponential tail bounds

The first lemma is a tail bound of a quadratic form of mean zero Gaussian variables proved by Laurent and Massart (2000).

Lemma II.10 (Laurent and Massart (2000))**.**

For $Z_1, ..., Z_n \overset{i.i.d.}{\sim} \mathbf{Norm}(0,1)$*, and* $\sigma_1, ..., \sigma_n \in \mathbb{R}^n_+$*, let* $W = \sum_{i=1}^n \sigma_i^2 Z_i^2$*, and*

$$a_1 = \sum_{i=1}^n \sigma_i^2,\ a_2^2 = \sum_{i=1}^n \sigma_i^4,\ a_\infty = \max_i \sigma_i^2 \ .$$

Then for any $x > 0$*,*

$$P\left(W > a_1 + 2a_2\sqrt{x} + 2a_\infty x^2\right) \leq \exp(-x)\ ,$$

$$P\left(W < a_1 - 2a_2\sqrt{x}\right) \leq \exp(-x)\ .$$

An immediate corollary is a tail bound on chi-square distribution.

Corollary II.11 (Laurent and Massart (2000))**.**

$$P(\sqrt{\chi_n^2/n} \geq 1 + \sqrt{2}t) \leq P(\chi_n^2/n \geq 1 + 2t + 2t^2) \leq \exp(-t^2 n)\ ,$$

$$P(\chi_n^2/n \leq 1 - 2t) \leq \exp(-t^2 n)\ .$$

The proof of Lemma II.10 is based on m.g.f. of chi-square distribution. The same idea can be used to prove a tail bound for non-central chi-square distribution.

Lemma II.12.

If $Z \sim \mathbf{Norm}(\mu, I_n)$*, where* $\mu \in \mathbb{R}^p$*,* $\|\mu\|_2 = c$ *and* $T = \|Z\|_2^2$*, then we have*

$$P\left(T \leq (c^2 + n) - 2\sqrt{2c^2 + n}x\right) \leq \exp(-x^2)\ ,$$

$$P\left(T \geq c^2 + n + 4\sqrt{c^2/2 + n/4}x + 2x^2\right) \leq \exp(-x^2)\ .$$

Proof. The m.g.f. of non-central chi-square distribution gives

$$\mathbb{E}[e^{uT}] = \exp\left(\frac{c^2u}{1-2u}\right)(1-2u)^{-n/2}, \quad u < 1/2 ,$$

so

$$\begin{aligned}\log \mathbb{E}[e^{uT}] &= \frac{c^2u}{1-2u} - \frac{n}{2}\log(1-2u) \\ &= c^2u\sum_{k=0}^{\infty}(2u)^k + \frac{n}{2}\sum_{k=1}^{\infty}\frac{(2u)^k}{k} \\ &= (c^2+n)u + (2c^2+n)u^2 + \sum_{k=3}^{\infty}\left(\frac{c^2}{2}+\frac{n}{2k}\right)(2u)^k .\end{aligned}$$

Consider the left tail. For any $t > 0$ and $u < 0$,

$$P(T \le c^2+n-t) \le \frac{\mathbb{E}[e^{uT}]}{e^{u(c^2+n-t)}} .$$

Since $u < 0$,

$$\log \mathbb{E}[e^{uT}] < (c^2+n)u + (2c^2+n)u^2 ,$$

and hence

$$\log P(T \le c^2+n-t) \le (2c^2+n)u^2 + ut ,$$

$$\log P(T \le c^2+n-t) \le \inf_{u<0}(2c^2+n)u^2 + ut = -\frac{t^2}{4(2c^2+n)} .$$

Let $t = 2\sqrt{2c^2+nx}$, and we prove the first part.

Now consider the right tail. For any $t > 0$ and $0 < u < 1/2$

$$P(T \ge c^2+n+t) \le \frac{\mathbb{E}[e^{uT}]}{e^{u(c^2+n+t)}} .$$

Since $0 < u < 1/2$,

$$\log \mathbb{E}[e^{uT}] \le (c^2+n)u + \left(\frac{c^2}{2}+\frac{n}{4}\right)\sum_{k=2}^{\infty}(2u)^k = (c^2+n)u + \left(\frac{c^2}{2}+\frac{n}{4}\right)\frac{4u^2}{1-2u} ,$$

and thus we have

$$\log P(T \ge c^2+n+t) \le \left(\frac{c^2}{2}+\frac{n}{4}\right)\frac{4u^2}{1-2u} - ut ,$$

$$\log P(T \geq c^2 + n + t) \leq \inf_{0<u<1/2} \left(\frac{c^2}{2} + \frac{n}{4}\right) \frac{4u^2}{1-2u} - ut \ .$$

Refer to Lemma 8. in Birgé and Massart (1998), we have

$$P(T \geq c^2 + n + 4\sqrt{c^2/2 + n/4}x + 2x^2) \leq \exp\{-x^2\} \ .$$

For completeness, proof is provided here. Denote $c' = c^2/2 + n/4$, and consider the function $h(u) = 4c'u^2/(1-2u) - ut$. We have

$$h(u) = \frac{c'}{1-2u} + \left(c' + \frac{t}{2}\right)(1-2u) - 2c' - \frac{t}{2} \ ,$$

$$\inf_{u\in(0,1/2)} h(u) = 2\sqrt{c'(c'+t/2)} - 2c' - t/2 \ .$$

If we let $t = 4\sqrt{c'}x + 2x^2$, then

$$\inf_{u\in(0,1/2)} h(u) = 2\sqrt{c'(c' + 2\sqrt{c'}x + x^2)} - 2c' - t/2 = 2\sqrt{c'}x - t/2 = -x^2$$

which complete the proof. □

The following lemma gives tail probability inequalities on Gaussian variable and extreme value of Gaussian variables.

Lemma II.13.

If $Z \in \mathbf{Norm}(0,1)$, then for any $t > 0$

$$P(Z > t) \leq \exp(-t^2/2) \ .$$

If $Z_1, ..., Z_p \sim \mathbf{Norm}(0,1)$, then for any $t > 0$:

$$P\left(\max_{j=1,\dots,p} Z_p > \sqrt{2\log p} + t\right) \leq P\left(\max_{j=1,\dots,p} Z_p > \sqrt{2\log p + t^2}\right) \leq \exp(-t^2/2) \ ,$$

$$P\left(\max_{j=1,\dots,p} |Z_p| > \sqrt{2\log p} + t\right) \leq P\left(\max_{j=1,\dots,p} |Z_p| > \sqrt{2\log p + t^2}\right) \leq 2\exp(-t^2/2) \ .$$

The following is a concentration inequality on the extreme singular values of Gaussian ensemble. See Vershynin (2010), where origin of such inequalities was discussed.

Lemma II.14 (Vershynin (2010))**.**

If $\mathbf{X} \in \mathbb{R}^{n\times p}$*, such that its rows are i.i.d. from* $\mathbf{Norm}(0, \Sigma_X)$*, where* $\Sigma_X \in \mathbb{R}^{p\times p}$*. Then with probability at least* $1 - \exp(-t/2)$*, we have*

$$\sigma_{\min}(\mathbf{X}) \geq \left(\sqrt{n} - \sqrt{p} - \sqrt{t}\right) \sqrt{\lambda_{\min}(\Sigma_X)}\ ,$$

and with probability at least $1 - \exp(-t/2)$*,*

$$\sigma_{\max}(\mathbf{X}) \leq \left(\sqrt{n} + \sqrt{p} + \sqrt{t}\right) \sqrt{\lambda_{\max}(\Sigma_X)}\ .$$

Proof. Vershynin (2010) gives proof for $\Sigma_X = I_D$. For general positive definite matrix Σ_X, we know that $\mathbf{X}\Sigma_X^{-1/2}$ is i.i.d. Gaussian ensemble. The proof is simple Corollary of the $\Sigma_X = I_D$ case. □

2.6.2 Eigenvalue and eigenvector perturbation

In this section, we discussed some classical results on eigenvalue and eigenvector perturbation. Singular value and vector perturbation is also closely related.

Lemma II.15 (Weyl's inequality)**.** *If* $X, E \in \mathbb{R}^{n\times m}$*, where* $n > m$*,* $\hat{X} = X + E$*. Let* $\sigma_j, \hat{\sigma}_j$ *be the* j*-th singular value of* $X, \hat{X}$*, respectively. Then*

$$|\sigma_j - \hat{\sigma_j}| \leq \sigma_{max}(E)\ .$$

Lemma II.16 (Wedin's sin(Θ) theorem, 1-dim, (Li 1998))**.** *If* $X = duv^T$ *and* $\hat{X} = duv^T + E$*. Let* $v, \hat{v}$ *be the top right singular vector of* $X, \hat{X}$*, respectively, and* $\theta = \angle(v, \tilde{v}) \in (0, \pi/2)$*. Then*

$$\sin\theta \leq \frac{2\sigma_{\max}(E)}{d}\ ,$$

which implies that

$$v^T\hat{v} = \cos(\theta) \geq 1 - \frac{2\sigma_{\max}(E)}{d} \ ,$$

$$\|v - \hat{v}\|_2 = 2\sin(\theta/2) \leq \sqrt{2}\sin\theta \ .$$

Lemma II.17 (Wedin's $\sin(\Theta)$ theorem, $D > 1$)**.** *If X be of rank D, and its SVD is given by $X = USV$, where $S = (\sigma_1, ..., \sigma_D)$. Also let $\hat{X} = X + E$, and $\hat{U}$, $\hat{V}$ be the top D left and right singular vectors of $\hat{X}$, and $\hat{\sigma}_d$ be the d-th singular values. Then*

$$\sigma_{\max}\left(\sin\Theta(V, \hat{V})\right) \leq \frac{2\sigma_{\max}(E)}{\sigma_D} \ ,$$

which implies that

$$\sigma_{\min}\left(V^T\hat{V}\right) \geq 1 - \frac{2\sigma_{\max}(E)}{\sigma_D} \ .$$

Proof. A typical Wedin's inequality (Li 1998) is

$$\max\left\{\left\|\sin\Theta(V, \hat{V})\right\|, \left\|\sin\Theta(U, \hat{U})\right\|\right\} \leq \frac{\max\{\|EV\|, \|U^TE\|\}}{\delta} \ ,$$

where δ is the "cross" singular value gap $(\sigma_D - \hat{\sigma}_{D+1}) \vee 0$ that involves both $X, X+E$. Any unitarily invariant norm can be used. If we use operator norm (i.e. the top singular value) in the inequality, then we get

$$\sin(\theta_D) \leq \frac{\sigma_{\max}(E)}{\delta} \ ,$$

because $\sigma_{\max}(EV) \leq \sigma_{\max}(E)\sigma_{\max}(V) = \sigma_{\max}(E)$ and $\sigma_{\max}\left(U^TE\right) \leq \sigma_{\max}(E)$ similarly.

If $\sigma_D < 2\sigma_{\max}(E)$, then $\frac{\sigma_{\max}(E)}{2\sigma_D} > 1$, so $\sin(\theta_D) \leq \frac{\sigma_{\max}(E)}{2\sigma_D}$ holds trivially.

If $\sigma_D \geq 2\sigma_{\max}(E)$, then note that i) $\sigma_{D+1} = 0$, ii) by Weyl's inequalty $\hat{\sigma}_{D+1} \leq \sigma_{D+1} + \sigma_{\max}(E) = \sigma_{\max}(E)$. We thus have $\sigma_D \leq 2\hat{\sigma}_{D+1}$, so $\sigma_D \geq 2(\sigma_D - \hat{\sigma}_{D+1}) = 2\delta$. Therefore, we still can conclude that $\sin(\theta_D) \leq \frac{2\sigma_{\max}(E)}{\sigma_D}$.

Once we prove $\sin(\theta_D) \leq \frac{2\sigma_{\max}(E)}{\sigma_D}$, $\sigma_{\min}\left(V^T\hat{V}\right) = \cos(\theta_D) \geq 1 - \sin(\theta_D) \geq 1 - \frac{2\sigma_{\max}(E)}{\sigma_D}$. □

Lemma II.18 (Davis-Kahan's sin(Θ) theorem, 1-dim)**.** *Let* $\Sigma = d^2 vv^T$, $\hat{\Sigma} = \Sigma + M$, *where* M *is a symmetric perturbation on* Σ. *Let* $\hat{v}$ *be the top eigenvector of* $\hat{\Sigma}$, *and* $\theta = \angle(v, \tilde{v})$. *Then*

$$\sin\theta \leq \frac{2\left\|M\right\|_{\text{op}}}{d^2} \ ,$$

which further implies that

$$v^T\hat{v} = \cos(\theta) \geq 1 - \frac{2\left\|M\right\|_{\text{op}}}{d^2} \ ,$$

$$\left\|v - \hat{v}\right\|_2 = 2\sin(\theta/2) \leq \sqrt{2}\sin\theta \ .$$

Lemma II.19 (Davis-Kahan's sin(Θ) theorem, $D > 1$)**.** *Let* Σ *is a symmetric positive definite matrix of rank* D, *with eigen-decomposition* $\Sigma = V\Lambda V^T$, *where* $\Lambda = \textbf{diag}(\lambda_1, ..., \lambda_D)$. *Also let* $\hat{\Sigma} = \Sigma + M$ *to be another SPD matrix, where* $\hat{V}$ *are the top* D *eigenvectors of* $\hat{\Sigma}$ *and* $\hat{\lambda}_d$ *are the d-th eigenvalue. Then*

$$\left\|\sin\Theta(V, \hat{V})\right\|_{\text{F}} \leq \frac{2\sqrt{D}\left\|M\right\|_{\text{op}}}{\lambda_D} \ .$$

Proof. A typical Davis-Kahan's inequality is (Li 1998)

$$\left\|\sin\Theta(V, \hat{V})\right\| \leq \frac{\left\|MV\right\|}{\delta} \ ,$$

where $\delta = (\lambda_D - \hat{\lambda}_{D+1}) \vee 0$, and any unitarily invariant norm can be used in it. If we use Frobenius norm, then

$$\left\|\sin\Theta(V, \hat{V})\right\|_{\text{F}} = \sqrt{\sum_{d=1}^{D}\sin(\theta_d)^2} \leq \frac{\left\|MV\right\|_{\text{F}}}{\delta} \leq \frac{\left\|V\right\|_{\text{F}}\left\|M\right\|_{\text{op}}}{\delta} = \frac{\sqrt{D}\left\|M\right\|_{\text{op}}}{\delta} \ .$$

If $\lambda_D \leq 2\left\|M\right\|_{\text{op}}$, then $\left\|\sin\Theta(V, \hat{V})\right\|_{\text{F}} \leq \sqrt{D} \leq 2\sqrt{D}\left\|M\right\|_{\text{op}}/\lambda_D$.

If $\lambda_D > 2\left\|M\right\|_{\text{op}}$, then note that i) $\lambda_{D+1} = 0$ ii) since $\Sigma, \hat{\Sigma}$ are both SPD matrices, so eigenvalues are also singular values, so by Weyl's inequality $\hat{\lambda}_{D+1} \leq \lambda_{D+1} + \left\|M\right\|_{\text{op}} = \left\|M\right\|_{\text{op}}$. Thus $\lambda_D > 2\hat{\lambda}_{D+1}$ which implies $\lambda_D < 2\delta$. Thus we still can conclude that $\left\|\sin\Theta(V, \hat{V})\right\|_{\text{F}} \leq 2\sqrt{D}\left\|M\right\|_{\text{op}}/\lambda_D$. Proof completed. □

CHAPTER III

Sliced Inverse Regression with Group Sparsity

3.1 Introduction

In the previous chapter, we discussed sparse PCA, which is a typical unsupervised dimension reduction method. Dimension reduction is also useful in supervised learning. If we have a response variable Y and high-dimensional predictors $X = (X_1, ..., X_p)$, and a linear subspace $\mathcal{S} \subseteq \mathbb{R}^D$ so that $Y \perp X \, \Big| \, P_{\mathcal{S}} X$, then $P_{\mathcal{S}} X$ provides all the information in X for explaining Y. The space $\mathcal{S}$ is called an efficient dimension reduction (EDR) space. Obviously EDR is not unique, and it is only meaningful to find a subspace that is as small as possible. Under mild condition (Cook 1996), the intersection of all the EDR spaces is also a EDR space, which is then the minimal dimension reduction space, and is called the central space. Methodology that estimates the central space is known as sufficient dimension reduction (SDR).

Among all SDR approaches, sliced inverse regression (Li 1991) is probably the most widely used one. Consider the multiple index model

$$Y = f(X\beta_1, ..., X\beta_D, \epsilon) \, , \tag{3.1}$$

where $\beta_1, ..., \beta_D$ are linearly independent and f is some unknown function. The noise ϵ is independent of X. The goal is to estimate the central space, which in this case is $\mathcal{S} = \mathbf{span}\langle \beta_1, ..., \beta_D \rangle$. Note that the actual coefficients $\beta_1, ..., \beta_D$ are not identifiable without knowing f.

Denote $\Sigma_X = \mathbf{cov}(X)$ and for simplicity, throughout the work we assume that $\mathbb{E}[X] = \mathbf{0}_p$, so we do not need to worry about centralization. The essential assumption behind SIR is the linearity condition, which is

$$\forall\, \xi \in \mathbb{R}^p,\ \exists\, c_1, ..., c_D,\ \text{s.t. } \mathbb{E}[\xi^T X | \beta_1^T X, ..., \beta_D^T X] = \sum_{d=1}^{D} c_d \beta_d^T X\ .$$

This is an assumption on the distribution of X. By assuming the linearity condition, one can prove that

$$m(y) := \mathbb{E}[X|Y = y] \in \Sigma_X \mathbf{span}\langle \beta_1, ..., \beta_D \rangle = \Sigma_X \mathcal{S},\ \forall y \in \mathbb{R}\ .$$

In SIR literature $m(y)$ is called the central curve. The property of central curve implies that

$$\mathbf{colspan}\langle \mathbf{cov}(m(Y)) \rangle \subseteq \Sigma_X \mathcal{S}\ .$$

If we further impose a coverage condition, that is,

$$\mathbf{colspan}\langle \mathbf{cov}(m(Y)) \rangle = \Sigma_X \mathcal{S}\ , \tag{3.2}$$

then $\mathcal{S}$ is the generalized eigenspace of $\mathbf{cov}(m(Y))$ with respect to Σ_X. In summary, linearity condition and coverage condition together make it possible to pose estimation of central space as a generalized eigenvalue problem (GEP). There are many other inverse regression methods as SAVE in Cook (2000) and MAVE in Xia et al. (2002), which are also based on similar thoughts. See Li (2007) for a more general formulation of inverse regression based SDR methods.

Sliced inverse regression provides a simple way to construct an empirical version of $\mathbf{cov}(m(Y))$. Specifically, assume $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y}, \boldsymbol{\epsilon} \in \mathbb{R}^n$ be i.i.d. samples from the multiple index model (3.1) and assume that $\mathbb{E}[X] = \mathbf{0}_p$. Divide the domain of Y

into H disjoint intervals (δ_h, δ_{h+1}), $h = 0, 1, ..., H-1$, and group the samples into H slices according to the value of Y, that is, the h-th slice has samples with index

$$\mathcal{G}_h = \{i : \mathbf{Y}_i \in (\delta_h, \delta_{h+1})\} \ .$$

The h-th slice mean of $\mathbf{X}$ is denoted by

$$\bar{\mathbf{X}}^{(h)} = \frac{1}{m_h} \sum_{i \in \mathcal{G}_h} \mathbf{X}_{i\cdot} \ ,$$

where $m_h = |\mathcal{G}_h|$ is the h-th slice size. Then SIR estimates $\mathbf{cov}(m(Y))$ by

$$\mathbf{M} := \sum_{h=1}^{H} \frac{m_h}{n} \bar{\mathbf{X}}^{(h)} (\bar{\mathbf{X}}^{(h)})^T \ .$$

In this work, instead of setting the division points $\delta_1, ..., \delta_{H-1}$, we slice the samples so that each slice has the same size. Let $n = Hm$, and $m_1 = ... = m_H = m$. Denote $\bar{\mathbf{X}}^{(*)} = (\bar{\mathbf{X}}^{(1)}, ..., \bar{\mathbf{X}}^{(H)})^T \in \mathbb{R}^{H \times p}$. Then $\mathbf{M} = \frac{1}{H} \bar{\mathbf{X}}^{(*)T} \bar{\mathbf{X}}^{(*)}$.

Either specifying division points or specifying slice sizes, one can estimate the central space by solving a GEP,

$$\mathbf{M}\mathbf{v}_d = \lambda_d \Sigma_X \mathbf{v}_d \ ,$$

where $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_D)$ satisfies $\mathbf{V}^T \Sigma_X \mathbf{V} = \mathbf{I}_D$. Usually Σ_X is unknown, in which case $\hat{\Sigma}_X = \mathbf{X}^T\mathbf{X}/n$ or some other estimator needs to be plugged in.

When $\Sigma_X = \mathbf{I}_p$, the GEP is replaced by a PCA problem. We have mentioned in the previous chapter that PCA is inconsistent when p diverges too fast with n; thus it is not surprising that SIR has the same issue. Indeed, Lin et al. (2015) shows under their assumptions that SIR is inconsistent when $p/n \not\to 0$. Thus, just like PCA, in order for SIR to be useful, some regularization is necessary and sparsity regularization can be used to both stabilize the estimation and improve interpretability.

There are many works that integrate sparsity into SIR. Some of them use penalization to get sparse estimators; see for example Zhong et al. (2005), Li and Nachtsheim

(2006), Li (2007), Li and Yin (2008), Chen et al. (2010), Yu et al. (2013), Tan et al. (2017), Lin et al. (2018). Others use thresholding to pre-screen the variables; see for example Jiang and Liu (2014), Yin and Hilafu (2015), Lin et al. (2015). Most of these works discuss sparse SIR algorithms without providing any statistical properties. Jiang and Liu (2014), Chen et al. (2010), Lin et al. (2015) provides asymptotic theories, which are not particularly useful when dimension is high. Recently Tan et al. (2017) and Lin et al. (2018) both propose algorithms that numerically work well and also provide theoretical results that are non-asymptotic. Basically, the distance between the true SDR space and the estimated space is often proved to be of rate $\frac{s \log p}{n\lambda}$, where λ is some model parameter that measures signal magnitude. Lin et al. (2018) require $n = o(\sqrt{p})$, while Tan et al. (2017) require $n = O(s^2 \log p/\lambda)$, and we can improve the sample size requirement.

3.2 Sparse SIR methods

In this section, we introduce some interesting SSIR methodologies. Similar to SPCA, Sparse SIR (SSIR) can also be achieved by either regularization or thresholding. The first method is so called "natural estimator" proposed by Tan et al. (2017). Secondly, we discuss an SSIR method based on Lasso proposed by Lin et al. (2018). Both methods use some penalization to induce sparsity. In addition, a third SSIR method based on thresholding is introduced, which has been proposed by Lin et al. (2015) but extended here. Finally, we introduce the "refinement" that has been utilized to improve the regularized estimators in Tan et al. (2017), which can also be applied to thresholded estimators.

3.2.1 SSIR using SPCA-SDP

Recall that the dimension reduction space can be obtained by solving the following generalized eigen-system

$$\mathbf{M}\mathbf{v}_d = \lambda_d \Sigma_X \mathbf{v}_d \ , \quad \mathbf{v}_d \Sigma_X \mathbf{v}_{d'} = \mathbf{1}\{d = d'\} \ .$$

where $\mathbf{M} = \bar{\mathbf{X}}^{(*)T}\bar{\mathbf{X}}^{(*)}/H$ is the matrix to be decomposed, Σ_X is the kernel, and $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_D)$ is orthonormal w.r.t. the kernel, $\mathbf{V}^T\Sigma_X\mathbf{V} = \mathbf{I}_D$. When Σ_X is unknown, it can be replaced by $\hat{\Sigma}_X = \mathbf{X}^T\mathbf{X}/n$. To get a sparse estimator of $\mathbf{V}$, one can formulate the GEP into an optimization problem with a sparsity penalty. Depending on the chosen optimization formulation and sparsity penalty, there can be different sparse GEP(SGEP) algorithms. One example is the “natural estimator” in Tan et al. (2017), which is a generalization of SPCA-SDP (Algorithm 3).

For simplicity, we assume Σ_X, Σ_X^{-1} are known; in case they are unknown, we can always replace them with suitable estimates. Consider the following optimization problem

$$\begin{aligned} \underset{V \in \mathbb{R}^{p \times D}}{\textbf{minimize:}} \quad & -\textbf{tr}\left(V^T \mathbf{M} V\right) + \rho \left\|V\right\|_{1,2} \\ \textbf{subject to:} \quad & V^T \Sigma_X V = \mathbf{I}_D \ . \end{aligned}$$

Notice the similarity between the above optimization criterion and ScotLass, except that instead of ℓ_1 norm constraint, we use group norm penalty. Like ScotLass, it cannot be solved efficiently, and needs to be relaxed.

The main difficulty of SGEP is the constraint $V^T\Sigma_X V = \mathbf{I}_D$, which is more complex than the orthonormal constraint $V^T V = \mathbf{I}_D$ in SPCA. As the result, many efficient relaxations that work for SPCA no longer work for SGEP. The reason is that the coordinate system under which V is sparse is different from the coordinate system where V is orthonormal (there is a linear transformation with $\Sigma_X^{1/2}$ in between).

Tan et al. (2017) use the SDP relaxation. Instead of optimizing the generalized eigenvectors V, they optimize over $F = VV^T$, which should also be sparse

$$\begin{aligned} \underset{F\in\mathbb{R}^{p\times p}}{\textbf{minimize:}} \quad & -\,\mathbf{tr}\,(\mathbf{M}F) + \rho\,\|F\|_1 \\ \textbf{subject to:} \quad & \left\|\Sigma_X^{1/2}F\Sigma_X^{1/2}\right\|_{\text{op}} \le 1,\ \left\|\Sigma_X^{1/2}F\Sigma_X^{1/2}\right\|_{*} \le D\ . \end{aligned}$$

To solve the above algorithm, ADMM can be used. The above optimization is equivalent to

$$\begin{aligned} \underset{F,G\in\mathbb{R}^{p\times p}}{\textbf{minimize:}} \quad & -\,\mathbf{tr}\,(\mathbf{M}F) + \rho\,\|F\|_1 + \infty\mathbf{1}\{\|G\|_{\text{op}} > 1\} + \infty\mathbf{1}\{\|G\|_{*} > D\} \\ \textbf{subject to:} \quad & \Sigma_X^{1/2}F\Sigma_X^{1/2} - G = 0\ , \end{aligned}$$

and the augmented Lagrangian is

$$\begin{aligned} \mathcal{L}_\eta(F,G,H) = {} & -\mathbf{tr}\,(\mathbf{M}F) + \rho\,\|F\|_1 + \infty\mathbf{1}\{\|G\|_{\text{op}} > 1\} + \infty\mathbf{1}\{\|G\|_{*} > D\} \\ & + \langle H, \Sigma_X^{1/2}F\Sigma_X^{1/2} - G\rangle + \frac{\eta}{2}\left\|\Sigma_X^{1/2}F\Sigma_X^{1/2} - G\right\|_{\text{F}}^2\ . \end{aligned}$$

The ADMM scheme is then iteratively updating F, G, H as

$$\begin{aligned} F^{(t+1)} &= \arg\min \mathcal{L}_\eta(F^{(t)}, G^{(t)}, H^{(t)})\ , \\ G^{(t+1)} &= \arg\min \mathcal{L}_\eta(F^{(t+1)}, G^{(t)}, H^{(t)})\ . \\ H^{(t+1)} &= H^{(t)} + \eta(\Sigma_X^{1/2}F^{(t+1)}\Sigma_X^{1/2} - G^{(t+1)})\ . \end{aligned} \tag{3.3}$$

When $\Sigma_X = \mathbf{I}_p$, the algorithm is SPCA-SDP; when $\Sigma_X \neq \mathbf{I}_p$, solving (3.3) is not easy. Tan et al. (2017) follow the proposal by Gao et al. (2017) where some general convex optimization solver called TFOCS is used to solve (3.3), but it is not as efficient as SPCA-SDP.

Another difficulty is that $\Sigma_X^{-1/2}$ needs to be computed in advance when solving (3.3). This is by itself a hard estimation problem when p is large. The author use pseudo-inverse, which empirically works quite well, but not well justified because pseudo-inverse may not be "close" to the true inverse. It is not easy to circumvent such complex computation or matrix inversion.

Algorithm 7 SSIR through SDP relaxation (SSIR-SDP)

```
Initialize $F^{(0)}, G^{(0)} \in \mathbb{R}^{p\times p}$, $H^{(0)} = \mathbf{0}_{p\times p}$.
For $k = 0, 1, 2, ...,$, repeat the following until convergence:
Solving $F^{(t+1)} = \arg\min \mathcal{L}_\eta(F^{(t)}, G^{(t)}, H^{(t)})$ using TFOCS solver.
$F^{(t+1)} = F^{(t+1)} + F^{(t+1)T}/2$
SVD: $\Sigma_X^{1/2} F^{(t+1)} \Sigma_X^{1/2} + H^{(t)}/\eta = \mathbf{LSR}$.
$G^{(t+1)} = \mathbf{LS'R}$, where $\mathbf{S'}$ is capped soft thresholding of $\mathbf{S}$, $\mathbf{s}'_j \leq 1$, $\sum_j \mathbf{s}'_j = D$.
$H^{(t+1)} = H^{(t)} + \eta(\Sigma_X^{1/2} F^{(t+1)} \Sigma_X^{1/2} - G^{(t+1)})$.
$\hat{\mathcal{S}}$ is the leading $D$-dimensional eigen-space of $\hat{F}$.
```

3.2.2 SSIR via Lasso

Another existing algorithm is called SIR-Lasso, proposed by Lin et al. (2018). Recall that the central space is given by

$$\mathcal{S} = \Sigma_X^{-1}\mathbf{colspan}\langle\mathbf{cov}(\mathbb{E}[X|Y])\rangle \ .$$

Let $\mathbf{J} \in \mathbb{R}^{n\times H}$ so that $\mathbf{J}_{ih} = 1$ if $\mathbf{Y}_i$ is in the h-th slice and $\mathbf{J}_{ih} = 0$ otherwise. Then $\bar{\mathbf{X}}^{(*)} = \frac{1}{m}\mathbf{J}^T\mathbf{X}$, so $\mathbf{M} = \frac{1}{m^2 H}\mathbf{X}^T\mathbf{J}\mathbf{J}^T\mathbf{X}$. Replace $\mathbf{cov}(\mathbb{E}[X|Y])$ by $\mathbf{M}$ and Σ_X by $\hat{\Sigma}_X = \mathbf{X}^T\mathbf{X}/n$. Consider the regular eigen-decomposition of $\mathbf{M}$,

$$\mathbf{M}\boldsymbol{\eta} = \boldsymbol{\eta}\boldsymbol{\Xi}, \quad \boldsymbol{\eta}^T\boldsymbol{\eta} = \mathbf{I}_D, \quad \Xi = \mathbf{diag}(\xi_1, ..., \xi_d) \ .$$

so that $\mathbf{colspan}\langle\boldsymbol{\eta}\rangle$ estimates $\mathbf{colspan}\langle\mathbf{cov}(\mathbb{E}[X|Y])\rangle$. Thus $\mathcal{S}$ can be estimated by $\mathbf{colspan}\langle\hat{\Sigma}_X^{-1}\boldsymbol{\eta}\rangle$. Notice that

$$\hat{\Sigma}_X^{-1}\boldsymbol{\eta} = \frac{1}{m}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{J}\mathbf{J}^T\mathbf{X}\boldsymbol{\eta}\boldsymbol{\Xi}^{-1} \ .$$

This is the least squares estimator of linear regression with predictor $\mathbf{X}$ and multivariate response $\tilde{\mathbf{y}} = \frac{1}{m}\mathbf{J}\mathbf{J}^T\mathbf{X}\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\Xi}}^{-1}$. To get a sparse estimator, we can then penalize the multivariate regression problem. Lin et al. (2018) use Lasso regression for each response variable separately. If we want the estimated eigenvectors to have common support, then the Lasso penalty can be replaced by a group Lasso penalty. The criterion then becomes

$$\min_{B\in\mathbb{R}^{p\times D}} \frac{1}{2n}\left\|\tilde{\mathbf{y}}-\mathbf{X}B\right\|_{\mathrm{F}}^2+\lambda\sum_{j=1}^{p}\left\|B_{j\cdot}\right\|_2 \ .$$

SSIR-Lasso circumvents estimation of Σ_X^{-1}, and to our knowledge it is the only SSIR algorithm in the literature that achieves this. Empirically, the performance is satisfactory when $D=1$. However when formulating SSIR into a multivariate linear regression, there are several approximations made in the process, which makes the justification less intuitive.

3.2.3 Thresholding-based SSIR

When $\Sigma_X=\mathbf{I}_p$, SSIR can be solved by using SPCA-DT or SPCA-AT on $\bar{\mathbf{X}}^{(*)}$ directly. The algorithms need to be adapted in order to work for general Σ_X. The idea of the adaption follows the work on sparse CCA proposed by Chen et al. (2013).

Let $\mathbf{M},\Sigma_X\in\mathbb{R}^{p\times p}$ be two symmetric matrices, and moreover $\mathbf{M}$ is positive semidefinite and Σ_X is positive definite. We want to solve the generalized eigen-system of $\mathbf{M}$ w.r.t. Σ_X, i.e.,

$$\mathbf{M}\mathbf{V}=\Sigma_X\mathbf{V}\mathbf{\Lambda},\quad \mathbf{V}^T\Sigma_X\mathbf{V}=\mathbf{I}_D \ .$$

If $\mathbf{M}$ is nearly rank D, i.e. the generalized eigenvalues beyond the D-th one are nearly 0, then we have

$$\Sigma_X^{-1}\mathbf{M}\Sigma_X^{-1}\approx\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{3.4}$$

$$\Sigma_X^{-1}\mathbf{M}\mathbf{V}\approx\mathbf{V}\mathbf{\Lambda} \tag{3.5}$$

Note that (3.4) is not an eigen-decomposition; the actual eigen-decomposition is $\Sigma_X^{-1/2}\mathbf{M}\Sigma_X^{-1/2} = (\Sigma_X^{1/2}\mathbf{V})\mathbf{\Lambda}(\Sigma_X^{1/2}\mathbf{V})^T$, and 3.4 can be derived from that.

If we assume that the row support of $\mathbf{V}$, $S = \{j : \mathbf{V}_{j\cdot} \neq 0\}$, is a small set. Then the r.h.s. of (3.4) has only a small non-zero block (rows and columns both in S), and the r.h.s. of (3.5) has row support S. This motivates the following SGEP algorithm, which is very similar to SPCA-AT. If we plug in $\mathbf{M} = \bar{\mathbf{X}}^{(*)T}\bar{\mathbf{X}}^{(*)}/H$, and some estimator $\hat{\Sigma}_X$ and $\hat{\Theta}_X$ of $\mathbf{cov}(X)$ and $\mathbf{cov}(X)^{-1}$, then the algorithm is an SSIR algorithm

Algorithm 8 SGEP via thresholding; also SSIR-DT and SSIR-AT

```
Input $\mathbf{M}, \hat{\Sigma}_X, \hat{\Theta}_X$, dimension $D$, threshold $\gamma_1, \gamma_2$
Calculate $W = \hat{\Theta}_X \mathbf{M} \hat{\Theta}_X$.
**Diagonal thresholding:** Using diagonal thresholding on $W$
    $$\hat{I} = W_{jj} \geq \gamma_1$$
**DT estimator:** Calculate the leading $D$ eigenvectors of $[\hat{\Sigma}_X]_{\hat{I}\hat{I}}^{-1/2}\mathbf{M}_{\hat{I}\hat{I}}[\hat{\Sigma}_X]_{\hat{I}\hat{I}}^{-1/2}$ to be $\hat{\mathbf{V}}_{raw}$.
    $\hat{\mathbf{V}}_{\hat{I}\cdot} = [\hat{\Sigma}_X]_{\hat{I}\hat{I}}^{-1/2}\mathbf{V}_{raw}$, and zero padding $\hat{\mathbf{V}}_{\hat{I}^c\cdot} = \mathbf{0}$.
Calculate $\tilde{W} = \hat{\Theta}_X \mathbf{M}\mathbf{V}$
**Extra thresholding:**
    $$\tilde{I}(\hat{I}, \gamma_2) = \{j \notin \hat{I} : \left\|\tilde{W}_{j\cdot}\right\|_2 > \gamma_2\} \cup \hat{I}$$
**Two stage estimator:** Calculate the leading $D$ eigenvectors of $[\hat{\Sigma}_X]_{\tilde{I}\tilde{I}}^{-1/2}\mathbf{M}_{\tilde{I}\tilde{I}}[\hat{\Sigma}_X]_{\tilde{I}\tilde{I}}^{-1/2}$ to be
    $\tilde{\mathbf{V}}_{raw}$. $\tilde{\mathbf{V}}_{\tilde{I}\cdot} = [\hat{\Sigma}_X]_{\tilde{I}\tilde{I}}^{-1/2}\mathbf{V}_{raw}$, and zero padding $\tilde{\mathbf{V}}_{\tilde{I}^c\cdot} = \mathbf{0}$.
$\hat{\mathbf{V}}$ is the SSIR-DT estimator and $\tilde{\mathbf{V}}$ is the SSIR-AT estimator.
```

3.2.4 Idea of refinement

In this section, we introduce a refinement technique that can be applied to any estimated generalized eigenvectors. This technique was inherited from the Sparse CCA algorithm proposed by Gao et al. (2017); Tan et al. (2017) apply it on the "natural estimator" which has been discussed in section 3.2.1. In some sense, the fundamental idea of this refinement is similar to SSIR-Lasso. The main difference is that in SSIR-Lasso, regular eigenvectors of $\mathbf{cov}(\mathbb{E}[X|Y])$ is refined instead of the generalized eigenvectors.

Recall that

$$\mathbb{E}[\mathbf{cov}(X|Y)]\mathbf{V} = \Sigma_X \mathbf{V}\mathbf{\Lambda} \ ,$$

and hence

$$\mathbb{E}[X\mathbb{E}[X|Y]^T]\mathbf{V} = \mathbb{E}[X(\mathbf{V}^T m(Y))^T] = \mathbb{E}[XX^T]\mathbf{V}\mathbf{\Lambda} \ .$$

Therefore, $\mathbf{V}\mathbf{\Lambda}$ solves the multivariate linear regression where $\mathbf{V}^T m(Y)$ is the multivariate response and X is the predictor. The response has $\mathbf{V}$ in it, and thus an initial estimator $\hat{\mathbf{V}}$ needs to be plugged in.

Let $\mathcal{J} \in \mathbb{R}^{n\times n}$ so that $\mathcal{J}_{i,j} = 1/m_h$ if $\mathbf{Y}_i, \mathbf{Y}_j$ are both in the h-th slice, and $\mathcal{J}_{i,j} = 0$ otherwise. Then the i-th row of $\mathcal{J}\mathbf{X}$ is $\bar{\mathbf{X}}^{(h)}$, so $\mathcal{J}\mathbf{X}$ are approximately samples of $\mathbb{E}[X|Y]$. Thus

$$\min_{U\in\mathbb{R}^{p\times D}} \frac{1}{n} \left\| \mathcal{J}\mathbf{X}\hat{\mathbf{V}} - \mathbf{X}U \right\|_{\mathrm{F}}^2 + \lambda \left\| U \right\|_{1,2}$$

gives a sparse estimator of $\mathbf{V}\mathbf{\Lambda}$. Note that $\mathbf{V}\mathbf{\Lambda}$ and $\mathbf{V}$ have the same column span, and the same row sparsity. Therefore the optimal solution of the criterion above also estimates central space.

If the initial estimator is the “natural estimator”, then we get the “refined SSIR” estimator in Tan et al. (2017). However we can also use thresholded SSIR as the initial estimator. Simulation results show that the refinement can drastically improve the thresholded estimator.

3.3 Statistical property

In this section, we provide some non-asymptotic theory on the statistical consistency of thresholded SSIR estimators when $D = 1$ and $\Sigma_X = \mathbf{I}_p$.

In some early works of SIR, theories are based on decent estimators of the covariance of central curve $\mathbf{cov}(\mathbb{E}[X|Y])$. If sample size is sufficiently large, we can have both sufficient number H of slices and also slices of large enough size m, so that $\mathbf{M}$ is close to $\mathbf{cov}(\mathbb{E}[X|Y])$ indeed. However, this never occurs in high dimensional setting where sample size is limited. In this section, we will show that the performance of SIR does not rely on $\mathbf{M}$ being close to $\mathbf{cov}(\mathbb{E}[X|Y])$ either; what is important is instead that slice means $\bar{\mathbf{X}}^{(1)}, ..., \bar{\mathbf{X}}^{(H)}$ approximately span the same space as the column span of $\mathbf{cov}(\mathbb{E}[X|Y])$. This can be achieved as long as the slice size m is large; slice number H does not have to be large.

3.3.1 Special case: Gaussian linear model

To reinforce the understanding, we start with a simple model. Assume Gaussian linear model with $\Sigma_X = \mathbf{I}_p$, i.e.,

$$Y = f(X\beta, \epsilon) = X^T\beta + \epsilon \ . \tag{3.6}$$

The inverse model can be written as

$$X = Y\mathbf{c} + U \ ,$$

where $Y \sim \mathbf{Norm}(0, \|\beta\|_2^2 + \sigma^2)$, $\mathbf{c} = \frac{\beta}{\|\beta\|_2^2+\sigma^2}$, $U \sim \mathbf{Norm}(\mathbf{0}_p, \Sigma_U)$ where $\Sigma_U = \mathbf{I}_p - \frac{\beta\beta^T}{\|\beta\|_2^2+\sigma^2}$, and Y is independent of U. The samples can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \ \Leftrightarrow \ \mathbf{X} = \mathbf{Y}\mathbf{c}^T + \mathbf{U} \ .$$

Since Y and U are independent, the data can be equivalently generated by first sampling $\mathbf{Y}$ and $\mathbf{U}$ independently according to their marginal distributions, and then calculating $\mathbf{X}$ using the inverse model.

Recall that we have $n = Hm$ samples, and are dividing them into H slices evenly according to the order of $\mathbf{Y}_i$'s. We use $\bar{\mathbf{X}}^{(h)}$ to denote the h-th slice mean of X, and $\bar{\mathbf{X}}^{(*)}$ to denote the collection of the H slice means. Use the same notation rule for other random variables/vectors. Then we have the slice-level aggregated inverse model

$$\bar{\mathbf{X}}^{(*)} = \bar{\mathbf{Y}}^{(*)}\mathbf{c}^T + \bar{\mathbf{U}}^{(*)} \,. \tag{3.7}$$

Since we assume $\Sigma_X = \mathbf{I}_p$ and $D = 1$, SSIR is just SPCA on $\bar{\mathbf{X}}^{(*)}$. If $\hat{\beta}^{DT}, \hat{\beta}^{AT} \in \mathbb{R}^p$ are the leading PC loadings of $\bar{\mathbf{X}}^{(*)}$ estimated by SPCA-DT and SPCA-AT, then $\mathbf{span}\langle\hat{\beta}^{DT}\rangle, \mathbf{span}\langle\hat{\beta}^{AT}\rangle$ are estimators of central space.

Note that the inverse model consists of two parts: a rank 1 signal $\bar{\mathbf{Y}}^{(*)}\mathbf{c}^T = \frac{\|\beta\|_2\bar{\mathbf{Y}}^{(*)}}{\|\beta\|_2^2+\sigma^2}\frac{\beta^T}{\|\beta\|_2}$, whose right singular vector is parallel to our target β; and a noise $\bar{\mathbf{U}}^{(*)}$, which in this special case is independent of the rank 1 signal. Thus the the inverse model is almost a single-spike model (2.28), except that now the left singular vector is random.

To study the statistical property of the SSIR-DT/SSIR-AT estimator under Gaussian linear model is essentially the same as studying the property of SPCA-DT/SPCA-AT. It can be easily done by conditioning on $\mathbf{Y}$ and applying Theorem II.7 directly. A crucial quantity in the theory of SPCA is the signal-to-noise ratio (of the inverse model, not the Gaussian linear model). Therefore, we need to first understand the magnitude of the signal part and noise part.

For the noise part, $\bar{\mathbf{U}}^{(*)}$ has i.i.d. rows following $\mathbf{Norm}(0, \Sigma_U/m)$. Thus the scale of the noise is $1/\sqrt{m}$. It is small when slices are large and it converges to 0 as m diverges. The number of slices does not influence this property.

On the other hand the signal part is of rank 1 even if $H = 2$ (if $H = 1$, then since we assume $\mathbb{E}[X] = 0$, $\bar{\mathbf{X}}^{(*)}$ and $\bar{\mathbf{Y}}^{(*)}$ are both close to 0, so the rank 1 signal degenerates to rank 0). The scale of the signal depends on $\left\|\bar{\mathbf{Y}}^{(*)}\right\|_2$, whose distribution is described as follows.

Definition III.1. *For H, m, let $n = Hm$ and suppose $Z_1, ..., Z_n \overset{i.i.d.}{\sim} \mathbf{Norm}(0, 1)$.*

Let the ordered statistics be $Z_{(1)}, ..., Z_{(n)}$ and

$$\bar{Z}^{(h)} = \frac{1}{m} \sum_{j=(h-1)m+1}^{hm} Z_{(j)} \ .$$

$$W^2 = \frac{1}{H} \sum_{h=1}^{H} \bar{Z}^{(h)2} \ .$$

Then $\Omega(H, m)$ denotes the distribution of W.

The distribution $\Omega(H, m)$ does not have a simple analytical form, but as long as we have more than two slices, the distribution is bounded away from 0. See remark below for some discussion. Obviously, with only two slices, $\mathbf{M}$ is not close to $\mathbf{cov}(m(Y))$ no matter how big m is; yet according to the theory of SPCA, SSIR-DT and SSIR-AT can both be consistent.

Remark III.2. *For fixed H, as $m \to \infty$, we have*

$$\Omega(H, m) \xrightarrow{d} \sqrt{\frac{1}{H} \sum_{h=1}^{H} \mathbb{E}\left[Z \;\middle|\; Z \in \left(\Phi^{-1}\left(\frac{h-1}{H}\right), \Phi^{-1}\left(\frac{h}{H}\right)\right)\right]^2} \ .$$

The right hand side is a constant (decided by H not m), and reach minimum when $H = 2$, where $\Omega(2, m) \xrightarrow{d} \mathbb{E}[Z|Z > 0]$. Thus, as long as $H \geq 2$, a variable from $\Omega(H, m)$ tends to be bounded away from 0. It can be easily proved that the rank 1 signal is of strength $\sigma_{\max}\left(\bar{\mathbf{Y}}^{(*)}c^T\right) \sim \frac{1}{\sqrt{1+(\sigma/\|\beta\|_2)^2}}\Omega(H, m)$, so it is also bounded away from 0 as long as $\sigma/\left\|\beta\right\|_2 = O(1)$.

Note that the variance of the slice means is smaller than the variance of the original samples. Therefore, as $n \to \infty$, $\Omega(m, H)$ should not exceed sample variance too much regardless of the trade-off between m and H. Therefore, intuitively, $\Omega(m, H)$ has the same rate as a constant.

The statistical property of SSIR-DT/AT under Gaussian linear model when $X \sim \mathbf{Norm}(\mathbf{0}_p, \mathbf{I}_p)$ is provided in the following theorem.

Theorem III.3. *Suppose that the data are generated from Gaussian linear model,*

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \ ,$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ *such that* $\mathbf{X}_{ij} \overset{i.i.d.}{\sim} \mathbf{Norm}(0,1)$ *and* $\boldsymbol{\epsilon} \in \mathbb{R}^n$ *such that* $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathbf{Norm}(0,\sigma^2)$*. Without loss of generality let* $\|\beta\|_2 = 1$*. Assume* $n = Hm$*, where* $P(\Omega(H,m) < c_1) \leq \delta_1$*, and denote* $\rho^2 = \frac{c_1^2}{1+\sigma^2}$*. Denote* $I_0 = \{j : \beta_j \neq 0\}$ *and* $s = |I_0|$*.*

Let $I_1 = \{j : |\beta_j| > \kappa_1\}$*,* $I_2 = \{j : |\beta_j| > \kappa_2\}$*, where*

$$t_0 = \sqrt{\frac{\log p - \log \delta_2}{n}} \ ,$$

$$\kappa_1 = \frac{2}{\rho}\sqrt{1 + 2t_0 + 2t_0^2} \ ,$$

$$\kappa_2 = \kappa_1 \wedge \kappa_2', \quad \kappa_2' = \frac{2}{\rho}\frac{\frac{\sqrt{2\log p}+t_3}{\sqrt{n}} + \frac{1}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{s+1}+\sqrt{2\log p}+t_4}{\sqrt{n}}\right)^2}{\sqrt{1-s\kappa_1^2} - \frac{2}{\rho\sqrt{m}} - \frac{1}{\rho}\frac{3\sqrt{s}+2t_1+t_2}{\sqrt{n}}} \ .$$

Assume that we can find constants $0 < \delta_2 < 1$ *and* $t_1, t_2, t_3, t_4 > 0$*, such that* $\kappa_2 > 0$*,* $\sqrt{s}\kappa_1 < 1$*. Then use SPCA-DT/AT on* $\bar{\mathbf{X}}^{(*)}$ *with the following thresholds*

$$\gamma_1 = (1 + 2t_0 + 2t_0^2) \ ,$$

$$\gamma_2 = \frac{1}{\rho}\frac{\sqrt{2\log p} + t_3}{\sqrt{n}} + \frac{(\sqrt{n} + \sqrt{s+1} + \sqrt{2\log p} + t_4)^2}{n} \ ,$$

we have

1. *with probability at least* $1 - \delta_1 - \delta_2$*, the DT-estimated set* $\hat{I}$ *satisfies*

$$I_1 \subseteq \hat{I} \subseteq I_0 \ ;$$

2. *with probability at least* $1 - \delta_1 - \delta_2 - 2\exp(-t_1^2/2) - \exp(-t_2^2/2)$ *the DT-estimated eigenvector* $\hat{\beta}^{DT}$ *satisfies*

$$\left\|\hat{\beta}^{DT} - \beta\right\|_2 \leq \sqrt{2s}\kappa_1 + \frac{4\sqrt{2}}{\rho\sqrt{1-s\kappa_1^2}}\left(\sqrt{\frac{s}{n}} + \frac{t_2}{\sqrt{n}}\right) + \frac{2\sqrt{2}}{\rho^2(1-s\kappa_1^2)}\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{s}{n}} + \frac{t_1}{\sqrt{n}}\right)^2 \ ,$$

where $\hat{\beta}^{DT}$ is the one between $\pm\hat{\beta}^{DT}$ that is closer to β;

3. *with probability at least* $1-\delta_1-\delta_2-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)-\exp(-t_4^2/2)$*, the AT-estimated set $\tilde{I}$ satisfies*

$$I_2 \subseteq \tilde{I} \subseteq I_0 \ ;$$

4. *with probability at least* $1-\delta_1-\delta_2-2\exp(-t_1^2/2)-\exp(-t_2^2/2)-2\exp(-t_3^2/2)-\exp(-t_4^2/2)$ *the AT-estimated eigenvector $\hat{\beta}^{AT}$ satisfies*

$$\begin{aligned}\left\|\hat{\beta}^{AT}-\beta\right\|_2 \le \sqrt{2s}\kappa_2 &+ \frac{4\sqrt{2}}{\rho\sqrt{1-s\kappa_2^2}}\left(\sqrt{\frac{s}{n}}+\frac{t_2}{\sqrt{n}}\right) \\ &+ \frac{2\sqrt{2}}{\rho^2(1-s\kappa_2^2)}\left(\sqrt{\frac{1}{m}}+\sqrt{\frac{s}{n}}+\frac{t_1}{\sqrt{n}}\right)^2 ,\end{aligned}$$

where $\hat{\beta}^{AT}$ is the one between $\pm\hat{\beta}^{AT}$ that is closer to β.

Proof. Denote $\bar{\mathbf{Z}}^{(*)} = \frac{1}{\sqrt{1+\sigma^2}}\bar{\mathbf{Y}}^{(*)}$. Consider the slice-level inverse model

$$\begin{aligned}\bar{\mathbf{X}}^{(*)} &= \bar{\mathbf{Y}}^{(*)}\mathbf{c}^T + \bar{\mathbf{U}}^{(*)} \\ &= \frac{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2}{\sqrt{H}\sqrt{1+\sigma^2}}\frac{\sqrt{H}\bar{\mathbf{Z}}^{(*)}}{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2}\beta^T + \bar{\mathbf{U}}^{(*)} .\end{aligned}$$

If we condition on $\mathbf{Y}$, then $\bar{\mathbf{Z}}^{(*)}$ is a constant vector. Conditionally, the inverse model becomes exactly the model in Theorem II.7. We have the following correspondence between model components and problem parameters:

$$\begin{aligned}\text{Theorem III.3} &\Leftarrow \text{Theorem II.7}\\ \sqrt{H}\frac{\bar{\mathbf{Z}}^{(*)}}{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2} &\Leftarrow \mathbf{u}\\ \beta &\Leftarrow \mathbf{v}\\ \mathbf{U} &\Leftarrow \boldsymbol{\epsilon}\\ H &\Leftarrow n\\ \frac{\mathbf{I}_p - \frac{\beta\beta^T}{1+\sigma^2}}{m} &\Leftarrow \Sigma_e\\ \frac{1}{m} &\Leftarrow \sigma_e\\ \frac{1}{\sqrt{1+\sigma^2}}\frac{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2}{\sqrt{H}} &\Leftarrow \sigma\end{aligned}$$

Denote the event $\mathcal{E} = \left\{\frac{1}{\sqrt{1+\sigma^2}}\frac{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2}{\sqrt{H}} \geq \rho\right\}$, and it is measurable w.r.t. $\mathbf{Y}$. Since $\frac{\left\|\bar{\mathbf{Z}}^{(*)}\right\|_2}{\sqrt{H}} \sim \Omega(H, m)$, by our assumption $P\left(\mathcal{E}^c\right) \leq \delta_1$, so ρ is a high probability lower bound of the magnitude of the low rank signal. Then we can use a conditional argument so that all the results in Theorem II.7 still hold, with δ_1 subtracted from the probability.

For example, by the first part of Theorem II.7, we have

$$P\left(I_1 \not\subseteq \hat{I} \text{ or } \hat{I} \not\subseteq I_0 \;\middle|\; \mathbf{Y}\right)\mathbf{1}\{\mathcal{E}\} \leq \delta_2 \,,$$

and hence

$$\begin{aligned}&P\left(I_1 \not\subseteq \hat{I} \text{ or } \hat{I} \not\subseteq I_0\right)\\ \leq &P\left(I_1 \not\subseteq \hat{I} \text{ or } \hat{I} \not\subseteq I_0, \mathcal{E}\right) + P(\mathcal{E}^c)\\ \leq &\int_{\mathcal{E}} P\left(I_1 \not\subseteq \hat{I} \text{ or } \hat{I} \not\subseteq I_0 | \mathbf{Y}\right) dP_{\mathbf{Y}} + P(\mathcal{E}^c)\\ \leq &\delta_2 + \delta_1 \,,\end{aligned}$$

$$P\left(I_1 \subseteq \hat{I} \subseteq I_0\right) \geq 1 - \delta_1 - \delta_2 \ .$$

The probability of the other three events can be bounded similarly. □

Corollary III.4. *Apply sparse SSIR-AT on independent Gaussian linear model. Assume $\sigma^2 / \|\beta\|_2^2 < c_1$, (H, m) satisfies that $P(\Omega(H, m) \leq c_2) \leq \delta$ for some constant c_1, c_2, so that $\frac{c_2^2}{1+c_1^2} m$ is large enough. Assume $\frac{s}{m}, \frac{s\sqrt{s}}{n}, \frac{s \log p}{n} \leq c$ for some small enough constant c. Then for some constant C, C', C'', c', c'', with probability at least $1 - \delta - C' \exp(-c's) - C'' p^{-c''}$,*

$$\left\|\hat{\beta}^{AT} - \beta\right\|_2 \leq C \left(\sqrt{\frac{s \log p}{n}} + \sqrt{\frac{s}{m}} + \frac{s\sqrt{s}}{n}\right) \ . \tag{3.8}$$

Moreover, simultaneously the estimated sparse set $\tilde{I}$ excludes all zero coordinates $\beta_k = 0$ and includes all coordinates so that $|\beta_k| \geq \tilde{C} \left(\frac{1}{m} + \frac{s}{n} + \sqrt{\frac{\log p}{n}}\right)$ for some $\tilde{C}$.

Remark III.5 (Comparison with the Lasso)**.** *There are various theoretical results for the Lasso. For example, in Wainwright (2009a), it is shown that the Lasso can recover the support of β if $n \succeq s \log(p-s)$ and $\beta_{\min} := \min\{|\beta_i| : \beta_i \neq 0\} \succeq \sqrt{\log p/n}$; moreover, $\left\|\hat{\beta} - \beta\right\|_\infty = O(\log p/n)$, so $\left\|\hat{\beta} - \beta\right\|_2 = O(s \log p/n)$. Comparing to that result, we have an extra term of rate $s\sqrt{s}/n$ which is hard to avoid with current proof technique. Another example is provided by Zhou (2009), where the author also proved that $\left\|\hat{\beta} - \beta\right\|_2^2 = O(s \log p/n)$ with a relaxed $\beta_{\min}$ condition. This is a possible direction of future work.*

3.3.2 General model, $D = 1$, $\Sigma_X = \mathbf{I}_p$

In this section we still assume that $X \sim \mathbf{Norm}(0, \mathbf{I}_p)$ and $D = 1$, but relax the Gaussian linear model into the single index model

$$Y = f(X^T\beta, \epsilon) \ ,$$

where we assume $X \sim \mathbf{Norm}(0, \mathbf{I}_p)$, and WLOG $\|\beta\|_2 = 1$. The inverse model is

$$X = m(Y) + U \ ,$$

where $m(Y) = \mathbb{E}[X|Y]$ is the central curve and $U = X - \mathbb{E}[X|Y]$ is the residual.

There are three major differences between the general case and the Gaussian linear model case. First, we still have $m(y) = c_y \beta$ for some c_y (because $m(y) \in \mathbf{span}\langle\beta\rangle$ by linearity condition, see section 3.1); however there is no easy analytical form of c_y in terms of y (under the Gaussian linear model, $c_y = \frac{y}{\|\beta\|_2^2 + \sigma^2}$). Second, U is not independent of Y, so when we use inverse model to generate data, we need to first generate $\mathbf{Y}_1, ..., \mathbf{Y}_n$, and then generate $\mathbf{U}_1, ..., \mathbf{U}_n$ accordingly. Third, The distribution of U is hard to characterize in general.

Denote $U|Y = y \sim F_{U,y}$. With $m(y)$ and $F_{U,y}$, the data $\mathbf{X}, \mathbf{Y}$ can be equivalently generated as follows

- Generate i.i.d. samples $\mathbf{Y}_1, ..., \mathbf{Y}_n$ from the distribution of $f(Z, \epsilon)$, where $Z \sim \mathbf{Norm}(0, 1)$.
- Calculate the central curve evaluated at the sample Y's, i.e. $m(\mathbf{Y}_1), ..., m(\mathbf{Y}_n)$.
- Generate the residuals $\mathbf{U}_1, ..., \mathbf{U}_n$, so that $\mathbf{U}_i$ follows $F_{U,\mathbf{Y}_i}$.
- $\mathbf{X}_i = m(\mathbf{Y}_i) + \mathbf{U}_i$.

We can still express the inverse model as single-spike model (2.28). To achieve this, denote the projection matrices $P_\beta = \beta\beta^T$ and $P_\beta^\perp = \mathbf{I}_p - P_\beta$, and notice that $m(Y)$ and U can be decomposed as

$$m(Y) = P_\beta \mathbb{E}[X|Y] = \mathbb{E}[\beta^T X|Y]\beta \ ,$$

$$U = X - m(Y) = P_\beta(X - m(Y)) + P_\beta^\perp X = (\beta^T U)\beta + P_\beta^\perp X \ .$$

Let $g(y) = \mathbb{E}[X^T\beta|Y = y] = m(y)^T\beta$. Also the conditional distribution $\beta^T U \Bigm| Y = y$ is denoted by $F_{E,y}$. This is also the conditional distribution of $Z - \mathbb{E}[Z|f(Z,\epsilon) = y]$ given $f(Z,\epsilon) = y$. We can then express the population inverse model into

$$X = (g(Y) + E_Y)\beta + P_\beta^\perp X \ ,$$

where given Y, E_Y is generated from $F_{E,Y}$. Again we get a low-rank signal plus a perturbation. $g(Y), E_Y$ are decided by $X\beta, \epsilon$, so they are independent of $P_\beta^\perp X$. Thus, to generate the sample, we can first generate Y, then generate E_Y according to Y, and then generate a multivariate Gaussian vector whose covariance matrix is $P_\beta^\perp$ to be "$P_\beta^\perp X$", and calculate X according to the inverse model. The sample version is

$$\mathbf{X}_{i\cdot} = (\mathbf{g}_i + \mathbf{E}_i)\beta + \mathbf{X}_{i\cdot} P_\beta^\perp \ ,$$

where $\mathbf{g}_i = m(\mathbf{Y}_i)$, and $\mathbf{E}_i|\mathbf{Y}_i$ follows $F_{E,\mathbf{Y}_i}$. Slice level aggregation gives

$$\bar{\mathbf{X}}^{(*)} = (\bar{\mathbf{g}}^{(*)} + \bar{\mathbf{E}}^{(*)})\beta + \bar{\mathbf{X}}^{(*)} P_\beta^\perp \ . \tag{3.9}$$

where $\bar{\mathbf{g}}^{(*)}$ and $\bar{\mathbf{E}}^{(*)}$ are within-slice means of $g(\mathbf{Y}_i)$'s and $\mathbf{E}_i$'s, respectively. The "noise" part $\bar{\mathbf{X}}^{(*)} P_\beta^\perp$ has i.i.d. rows following $\mathbf{Norm}(\mathbf{0}_p, P_\beta^\perp/m)$.

The statistical property of SSIR-DT/AT applied to single index model is similar to that applied to Gaussian linear model, and Theorem III.3 can be adapted with one change. Instead of conditioning on $\mathbf{Y}$, we now need to condition on $\mathbf{X}\beta, \boldsymbol{\epsilon}$. With the conditioning, $\bar{\mathbf{g}}^{(*)}, \bar{\mathbf{E}}^{(*)}$ are both constant vectors. Just like in GLM, we need to have a high-probability lower bound of the magnitude of the signal part in the inverse model, which in this case is decided by $\left\|\bar{\mathbf{g}}^{(*)} + \bar{\mathbf{E}}^{(*)}\right\|_2 / \sqrt{H}$. To achieve this, we make the following assumption on (f, ϵ, H, m).

Assumption III.6. *Let* $Z \sim \mathbf{Norm}(0,1)$*,* $g(y) = \mathbb{E}[Z \,\Big|\, f(Z,\epsilon) = y]$*, and let* $F_{E,y}$ *be the conditional distribution* $Z - \mathbb{E}[Z \,\Big|\, f(Z,\epsilon) = y] \,\Big|\, f(Z,\epsilon) = y$*. Let* $n = Hm$*. We say that* $(f, \epsilon, H, m) \in \mathcal{C}(c_1, c_2, p_1, p_2)$ *if the following two inequalities holds:*

- *With* $\mathbf{Y}_1, ..., \mathbf{Y}_n$ *being i.i.d. samples of* $f(Z,\epsilon)$*, calculate* $\mathbf{g}_i = g(\mathbf{Y}_{(i)})$*, divide them into* H *slices and denote the vector of within-slice means by* $\bar{\mathbf{g}}^{(*)} \in \mathbb{R}^H$*. Then*
$$P\left(\frac{\left\|\bar{\mathbf{g}}^{(*)}\right\|_2}{\sqrt{H}} < c_1\right) \leq p_1 \ .$$
- *For any* n *sorted numbers* $y_1 < y_2, ..., < y_n$*, let* $\mathbf{E}_i \sim F_{E,y_i}$*. Denote* $\bar{\mathbf{E}}^{(*)} \in \mathbb{R}^H$ *to be the within-slice means of* $\mathbf{E}_i$*'s. Then*
$$P\left(\frac{\left\|\bar{\mathbf{E}}^{(*)}\right\|_2}{\sqrt{H}} > \frac{c_2}{\sqrt{m}}\right) \leq p_2 \ .$$

Assumption III.6 is obviously posed for technical reason. It is hard to check whether the assumption holds for most (f, ϵ, H, m), but the assumption can be understood intuitively.

The first part of the assumption is essentially to assume that inverse regression curve is not "flat". We know that some functions f are problematic; for example if $f(x,\epsilon) = x^2 + \epsilon$, then $\mathbb{E}[Z|Z^2 + \epsilon] = 0$, so g is always 0. This assumption eliminates that.

The second part is made on "variance". There is no straightforward analysis, but generally speaking the more variant $f(Z,\epsilon)$ is from some deterministic function $f(Z)$ the larger $\left\|\bar{\mathbf{E}}^{(*)}\right\|_2$ is, so that the constants in the assumption is worse. Some special cases to help understanding: 1) if $f(Z,\epsilon) = f(Z)$, where f is strictly monotonic, then $\left\|\bar{\mathbf{E}}^{(*)}\right\|_2 = 0$; 2) if $f(Z,\epsilon) = Z + \epsilon$, where $\epsilon \sim \mathbf{Norm}(0, \sigma^2)$, then $\bar{\mathbf{E}}^{(*)}$ has entries i.i.d. following $\mathbf{Norm}(0, (1 - \frac{1}{1+\sigma^2})/m)$, so the smaller σ^2 is the better.

With the above assumptions we can state the following theorem on consistency of sparse SIR.

Theorem III.7. *Suppose that data are generated from single index model,*

$$\mathbf{Y}_i = f(\mathbf{X}_i^T \beta, \boldsymbol{\epsilon}_i) \ ,$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ *such that* $\mathbf{X}_{ij} \overset{i.i.d.}{\sim} \mathbf{Norm}(0,1)$. *Without loss of generality, let* $\|\beta\|_2 = 1$. *Assume that* $(H, m, f, \epsilon) \in \mathcal{C}(c_1, c_2, \delta_1, \delta_2)$ *defined in Assumption III.6, and denote* $\rho = c_1 - c_2/\sqrt{m}$. *Then all the results in Theorem III.3 holds.*

3.3.3 Discussions

We compare our theory with that in Tan et al. (2017). Their results generalize to $D > 1$ and $\Sigma_X \neq \mathbf{I}_p$; we only proved theory for $D = 1$ and $\Sigma_X = \mathbf{I}_p$, but potentially this can also be generalized. We also make different assumptions. In Tan et al. (2017), Y is categorical with H values, and F_{E,y_h} is $\mathbf{Norm}(\mathbf{0}, \Sigma_h)$; in our work, we assume $X \sim \mathbf{Norm}(\mathbf{0}, \mathbf{I}_p)$, but F_{E,y_h} can be some general distribution. Their assumption almost implies our assumption, except that they assume conditional distribution to be Gaussian while we assume marginal distribution to be Gaussian. Another difference is that they directly make the generalized eigenvalues of $\mathbf{cov}(\mathbb{E}[X|Y])$ w.r.t. $\mathbf{cov}(X)$ parameters, and assume that the minimal non-zero generalized eigenvalue is bounded away from 0. This is actually a form of coverage assumption, that is: $\Sigma_X^{-1}\mathbf{colspan}\langle m(Y_1), ..., m(Y_n)\rangle$ is not a subset of $\mathcal{S}$ but equals $\mathcal{S}$. If some generalized eigenvalue is close to zero, than the corresponding eigenvector will not be covered by that space. In our work, the assumption on $\Omega(H, m)$ in Theorem III.3 and part 1 of Assumption III.6 in Theorem III.7 are imposed for this purpose.

As for the error bounds of the final estimator, Tan et al. (2017) provided a lower bound

$$C\frac{s\log(p/s)}{n}, \text{ assuming } \frac{s\log(p/s)}{n} \leq c\ ,$$

and an upper bound

$$C\frac{s\log p}{n}, \text{ assuming } \frac{s^2\log p}{n} \leq c\ .$$

We only provide an upper bound. If we fix H as what they did, then our upper bound is

$$C\left(\frac{s\log p}{n}+\frac{s^3}{n^2}\right), \text{ assuming } \frac{s\sqrt{s}}{n}, \frac{s\log p}{n} \leq c\ .$$

Since $\frac{s\sqrt{s}}{n}, \frac{s\log p}{n} = O(\frac{s^2\log p}{n})$, our assumption on (n, p, s) is less demanding asymptotically. The extra term in the error, $\frac{s^3}{n^2}$, is not always of smaller rate than $\frac{s\log p}{n}$; however, if we do assume Tan's sample size requirement that $\frac{s^2\log p}{n} = O(1)$, then $\frac{s^3}{n^2} = \frac{s}{n} \times \frac{s^2}{n} = O(\frac{s}{n}) = O(\frac{s\log p}{n})$. Thus, our theory is slightly better than theirs asymptotically.

3.4 Simulation on SSIR-DT/AT

In this section, we run numerical study to compare different versions of thresholding based methods, to check whether our theoretical results are correct, and to see the performance of those methods in the problem regime for which we have not developed any theory ($D > 1$, $\Sigma \neq I_p$ and refinement).

3.4.1 Asymptotic rate of sample size

The purpose of the first batch of experiments is to see how sample size n scales with dimension p. We generate data from the following Y models:

- **Model I**: $D = 1$, $Y = X\beta_1 + \sin(X\beta) + \epsilon$, $\beta_1 = (b_s, \mathbf{0}_{p-s})$.
- **Model II**: $D = 1$, $Y = (X\beta_1)^3 + \epsilon$, $\beta_1 = (b_s, \mathbf{0}_{p-s})$.

- **Model III**: $D = 2$, $Y = (X\beta_1)\exp(X\beta_2) + \epsilon$, $(\beta_1, \beta_2) = \begin{pmatrix} b_s & \mathbf{0}_s \\ \mathbf{0}_s & b_s \\ \mathbf{0}_{p-2s} & \mathbf{0}_{p-2s} \end{pmatrix}$.

- **Model IV**: $D = 2$, $Y = (X\beta_1) + \exp(X\beta_2) + \epsilon$, $(\beta_1, \beta_2) = \begin{pmatrix} b_s & \mathbf{0}_s \\ \mathbf{0}_s & b_s \\ \mathbf{0}_{p-2s} & \mathbf{0}_{p-2s} \end{pmatrix}$.

- **Model V**: $D = 2$, $Y = (X\beta_1)(1+X\beta_1+X\beta_2)+\epsilon$, $(\beta_1, \beta_2) = \begin{pmatrix} b_s & \mathbf{0}_s \\ \mathbf{0}_s & b_s \\ \mathbf{0}_{p-2s} & \mathbf{0}_{p-2s} \end{pmatrix}$.

where $X \sim \mathbf{Norm}(\mathbf{0}, \Sigma_X)$, and we tested two different Σ_X: $\Sigma_1 = I_p$ and Σ_2 such that $[\Sigma_2]_{ij} = 0.3^{|i-j|}$. For the purpose of this experiment, we assume that Σ_X is known. We fix the variance of noise $\epsilon \sim \mathbf{Norm}(0, 0.3^2)$, $s = 30$, and $b_s = \kappa(0.8, 0.8^2, ..., 0.8^s)$, where κ is chosen so that $\|b_s\|_2 = 1$.

We ran four different methods based on thresholding: SSIR-DT, SSIR-AT, and SSIR-DT/AT with refinement (SSIR-DT-ref, SSIR-AT-ref). We use $H = 8D$ slices. All these methods have tuning parameter(s), and we only vary one of them. For DT, we set $\gamma_1 = (0.05, 0.1, ..., 1) \times \sqrt{\frac{1.5\log p + H/2}{n}}$; for AT, we fix $\gamma_1 = 0.5\sqrt{\frac{1.5\log p+H/2}{n}}$, and set $\gamma_2 = (0.2, 0.4, ..., 4) \times \log p/n$; for DT+ref, we fix $\gamma_1 = 0.3\sqrt{\frac{1.5\log p+H/2}{n}}$, and set the tuning parameter in the refinement by $\lambda = (0.1, ..., 2)\sqrt{\log p/n}$; for AT-ref, we fix $\gamma_1 = 0.5\sqrt{\frac{1.5\log p+H/2}{n}}$, $\gamma_2 = 2\log p/n$ and set $\lambda = (0.1, ..., 2)\sqrt{\log p/n}$. The scale of γ_1, γ_2 are set according to the theory, and the scale of λ is what appear in the literature. The constants are not carefully picked, and in fact there is not one good constant that works for all model settings.

When fixing one model setup ($Y \sim X$ model and Σ_X model), we vary $p = 150, 300, 600, 1200$ and a range of n. For each (n, p) combination, we generate 200 independent data sets, and run the 4 methods each under 20 tuning parameters. We calculate the average $\mathbf{Dist}_{ave}$ and take the minimum among the 20 tuning parameters, so that each method gives one minimal average error. We plot the error against raw sample size n and scaled sample size $n/\log p$ in Figure 2.1 to 2.10. Each figure has 8 panels, corresponding to 4 methods, and x-axis being raw or scaled n. Each panel has 4 curves, corresponding to 4 different p.

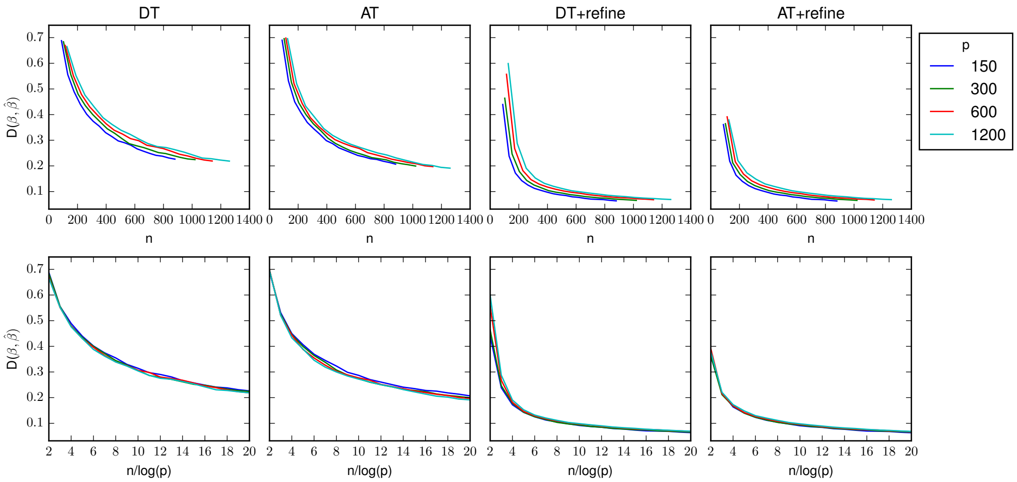


Figure 3.1: Model I, Σ_1

One can see that before scaling, the curves from larger p are above the curves from smaller p, but after scaling, either the curves overlap or the curves from smaller p shift above those curves from larger p. This means that with other model parameter fixed and p increasing, $n \asymp \log p$ is sufficient to keep the error not increasing.

Another observation is that refinement actually improves the results significantly. To see this better, we can rearrange the curves and plot together the curves from different methods but same model setups. See Figure 2.11 and 2.12.

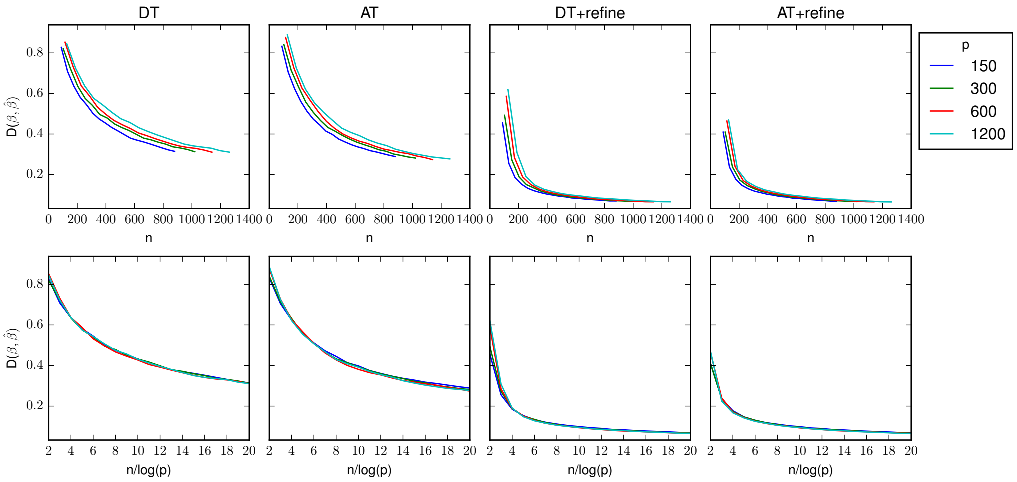


Figure 3.2: Model I, Σ_2

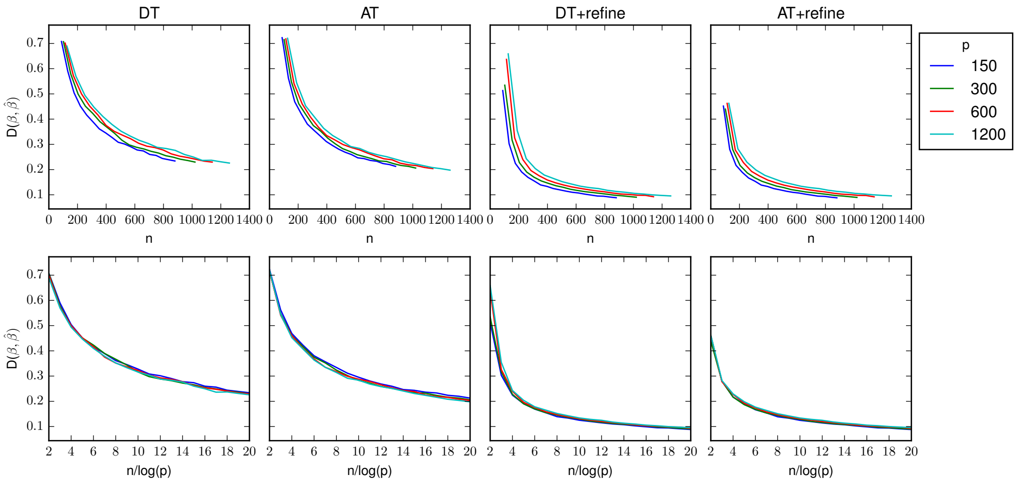


Figure 3.3: Model II, Σ_1

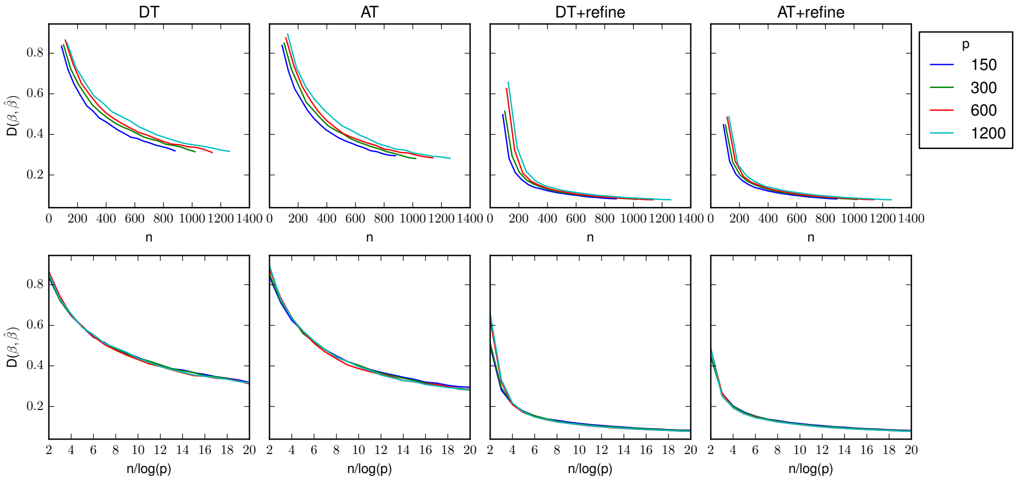


Figure 3.4: Model II, Σ_2

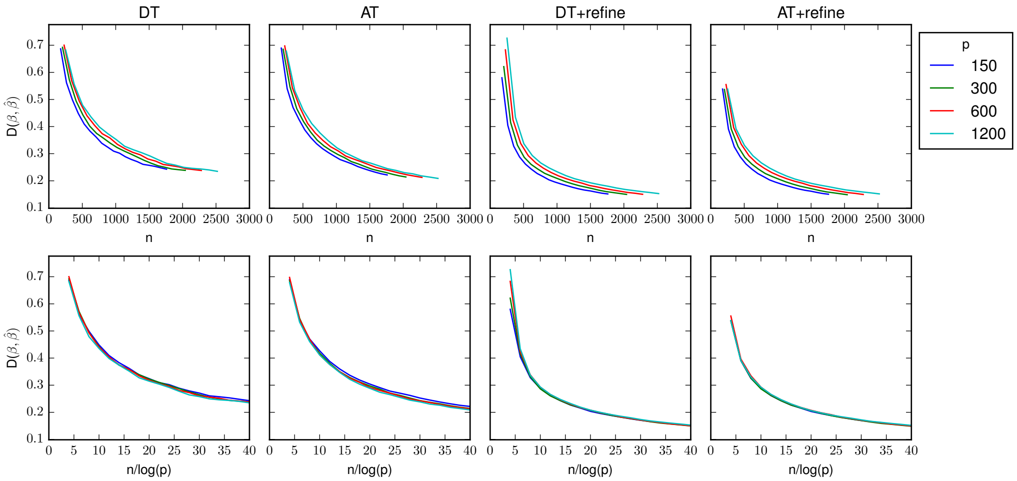


Figure 3.5: Model III, Σ_1

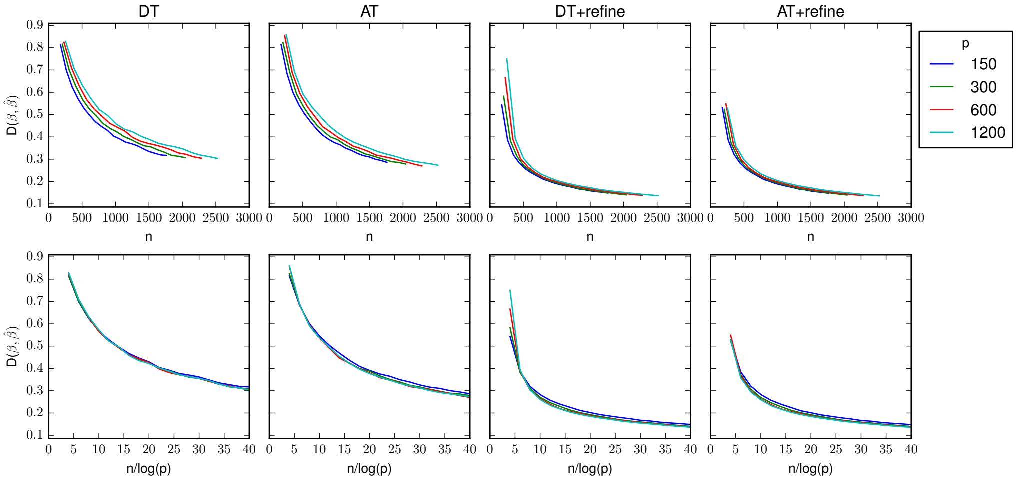


Figure 3.6: Model III, Σ_2

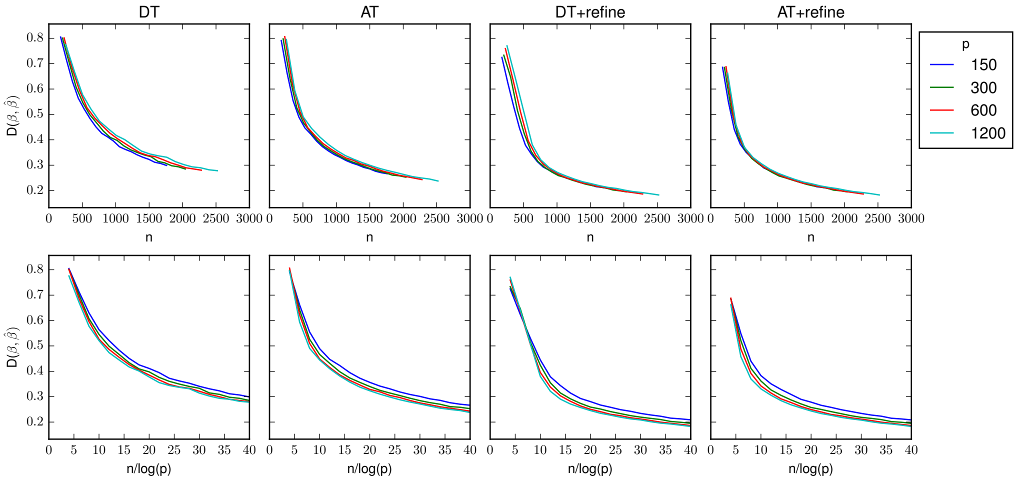


Figure 3.7: Model IV, Σ_1

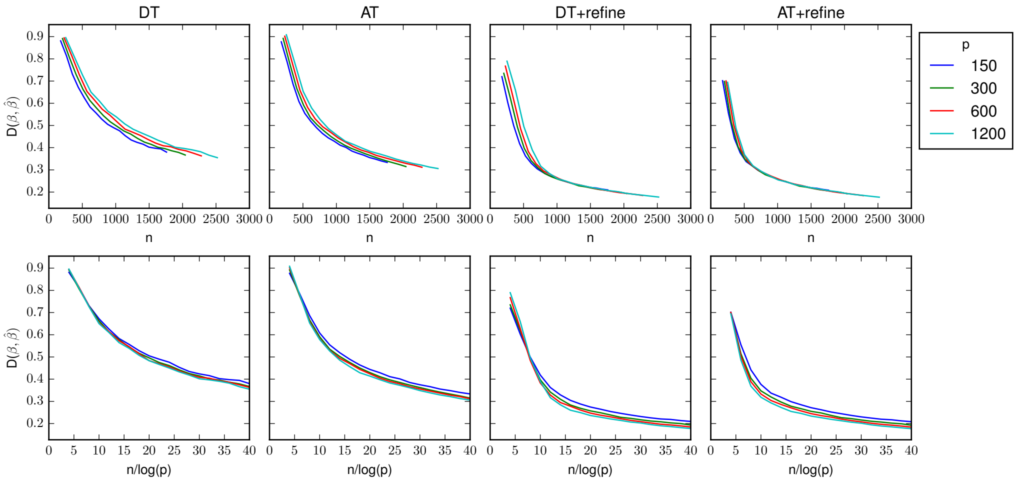


Figure 3.8: Model IV, Σ_2

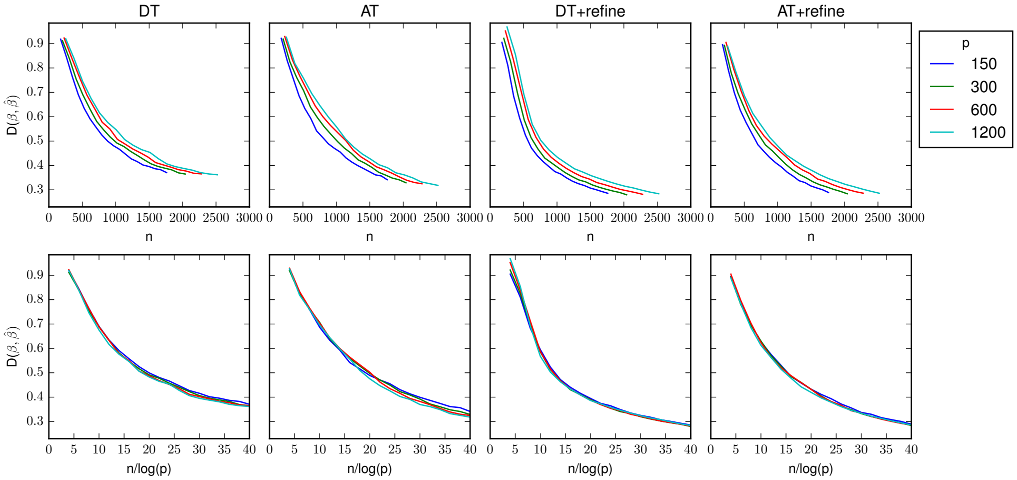


Figure 3.9: Model V, Σ_1

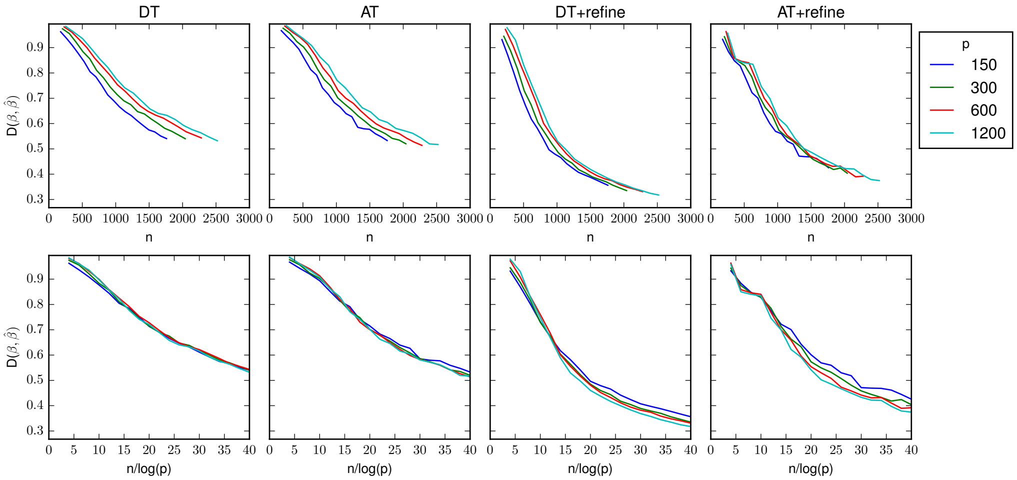


Figure 3.10: Model V, Σ_2

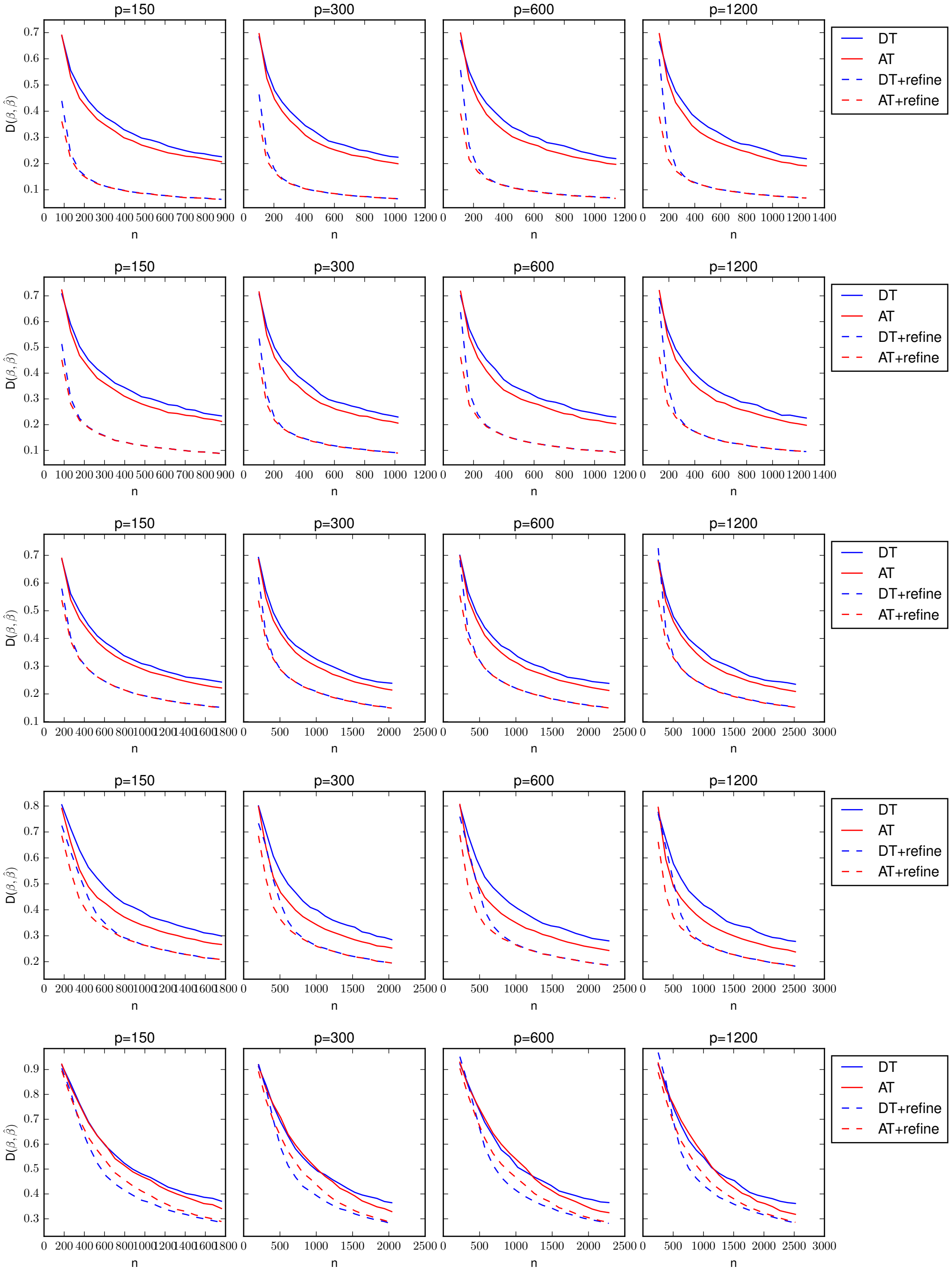


Figure 3.11: Rows corresponding to Model I to V, respectively; $\Sigma_X = \Sigma_1$

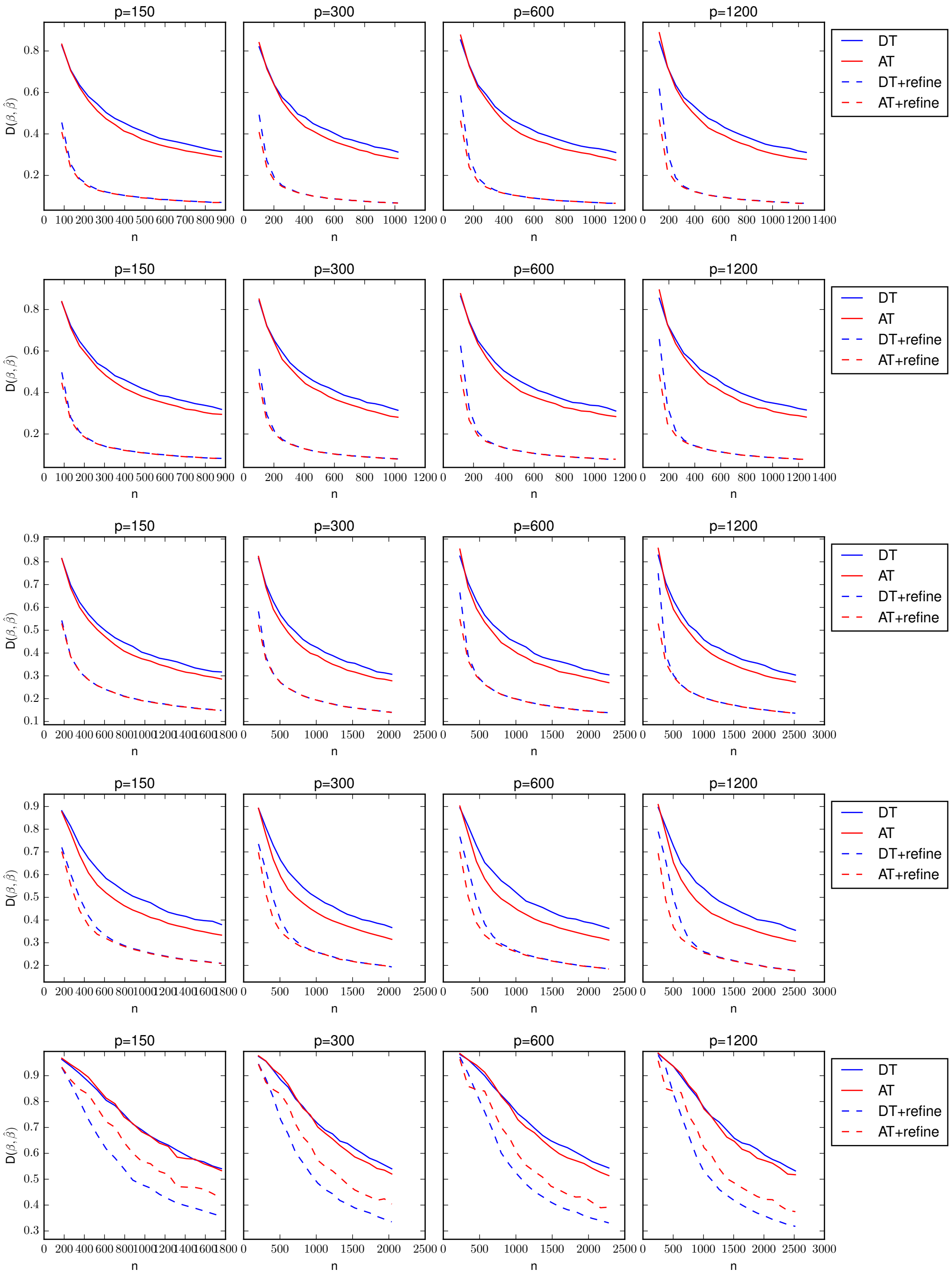


Figure 3.12: Rows corresponding to Model I to V, respectively; $\Sigma_X = \Sigma_2$

We have three observations. First, for all 5 models, improvement from refinement is obvious. Second, it seems that in many cases, the errors after refinement are not substantially affected by the initial estimators. Finally, the errors do not increase dramatically even when p increases rapidly, which means that thresholding effectively stabilizes the estimation.

3.4.2 Choice of slice number H

Another interesting question is how the number of slices in SIR affects the results. To investigate the impact of slice number H, we fix $p = 600$, and vary $n = 500, 700, 1000, 1400$ and for a specific n we then vary H as integer part of $(3, 4, 6, 8, 12, 16, 24, 32, 40, 48, 56, 64) \times n/500$. When tuning γ_1, we use the base $\sqrt{\frac{1.5 \log p + H/4}{n}}$ instead of $\sqrt{\frac{1.5 \log p + H/2}{n}}$, to mitigate the impact of large H on the range of tuning parameters.

The results are shown in Figure 2.13-2.17. We can see that having a lot of small slices is worse than having a few large slices. Look at the plots where x-axis is H/n (meaning that the results are aligned according to the slice size $m = n/H$), one can see that the optimal choices of H also get aligned, and under the tested settings (Y vs X models, Σ_X, and parameters like σ, p) usually it is better to have at least 50 samples per slice, and definitely not less than 25 samples per slice (because optimal H/n is always around 0.02 to 0.04), even if keeping that slice size means that we only have very few slices. In fact, in the literature, usually H is considered as fixed parameter and pick a small value in application. Here, we provide systematic experiments to justify that choice of small H (most of time in the range of 8-20), because the errors increase rapidly when over-slicing, but are stable when under-slicing.

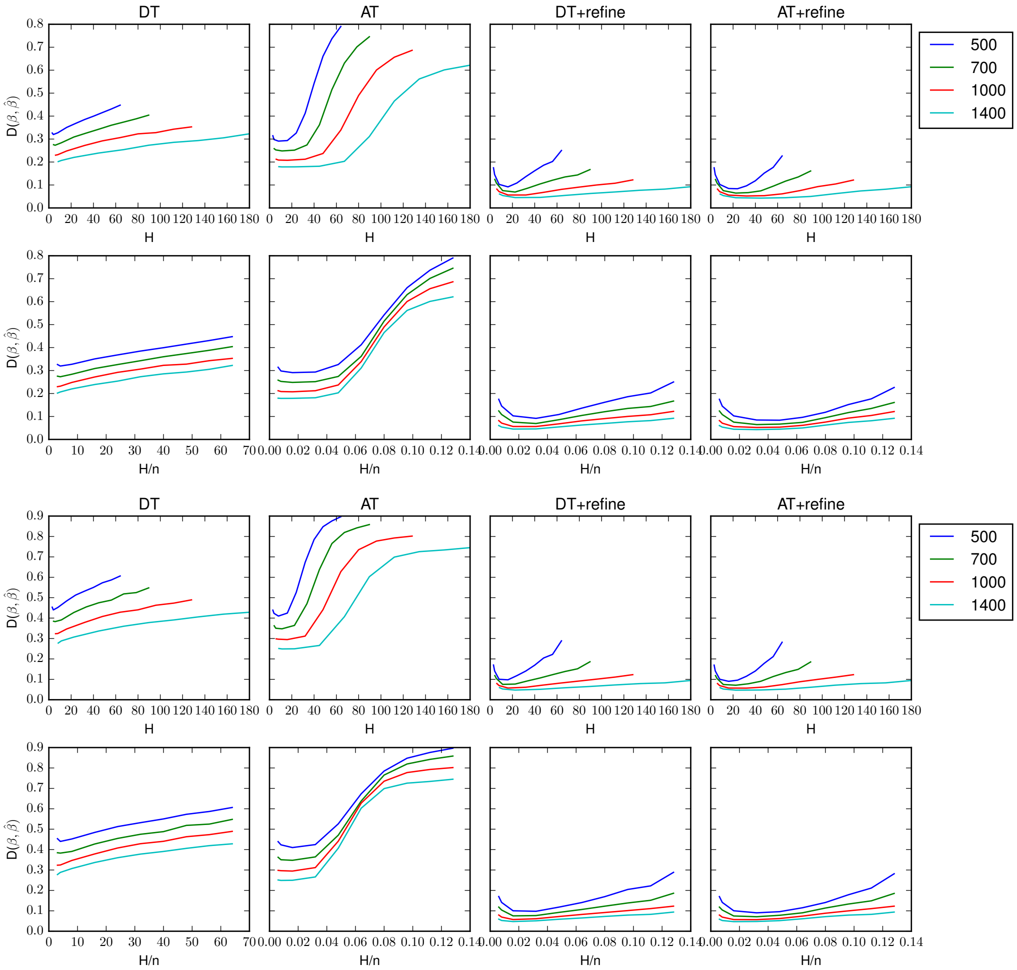


Figure 3.13: Model I, top two rows: Σ_1, bottom two rows: Σ_2.

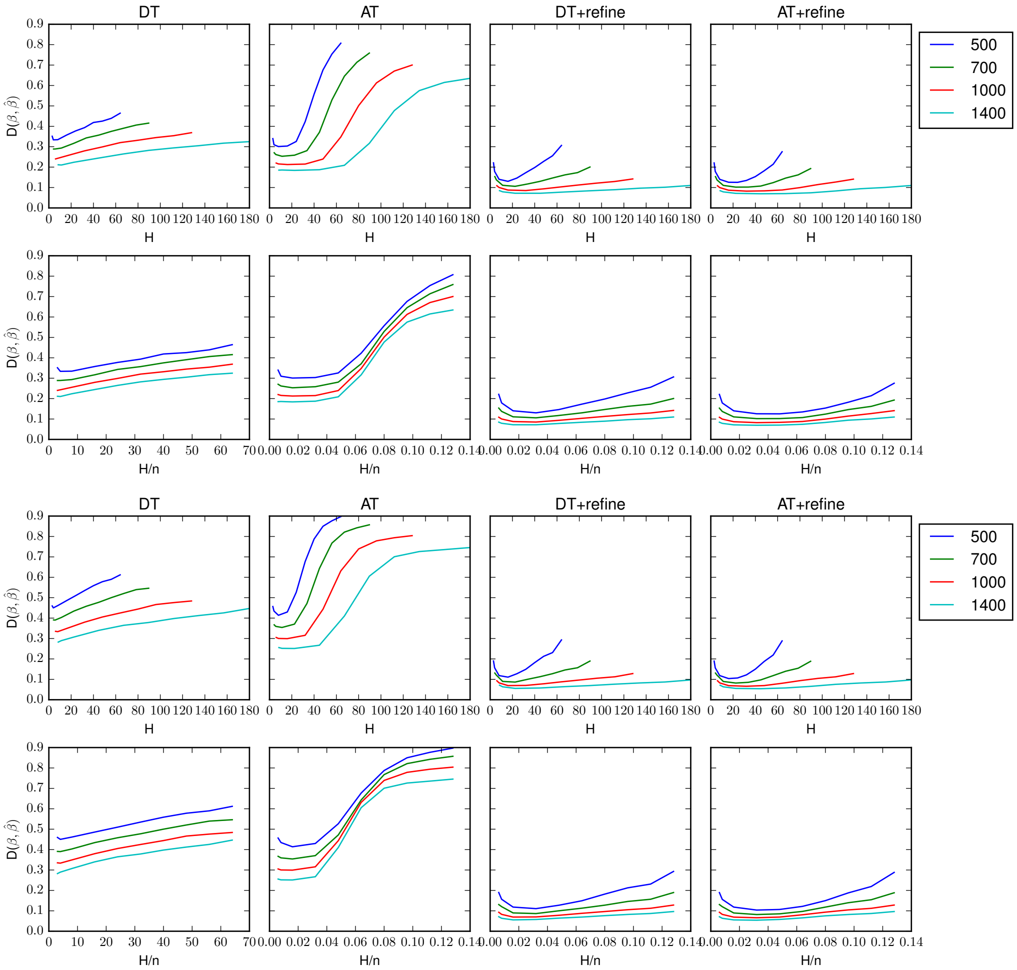


Figure 3.14: Model II, top two rows: Σ_1, bottom two rows: Σ_2.

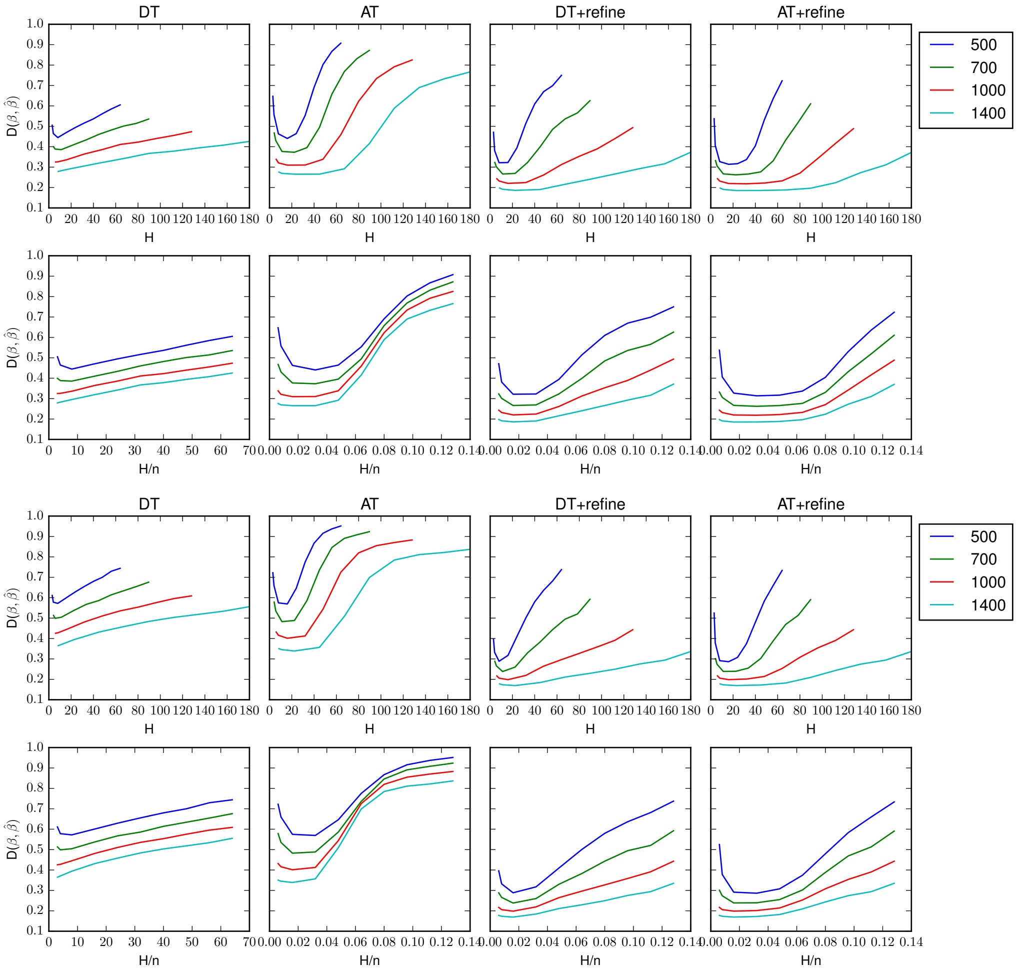


Figure 3.15: Model III, top two rows: Σ_1, bottom two rows: Σ_2.

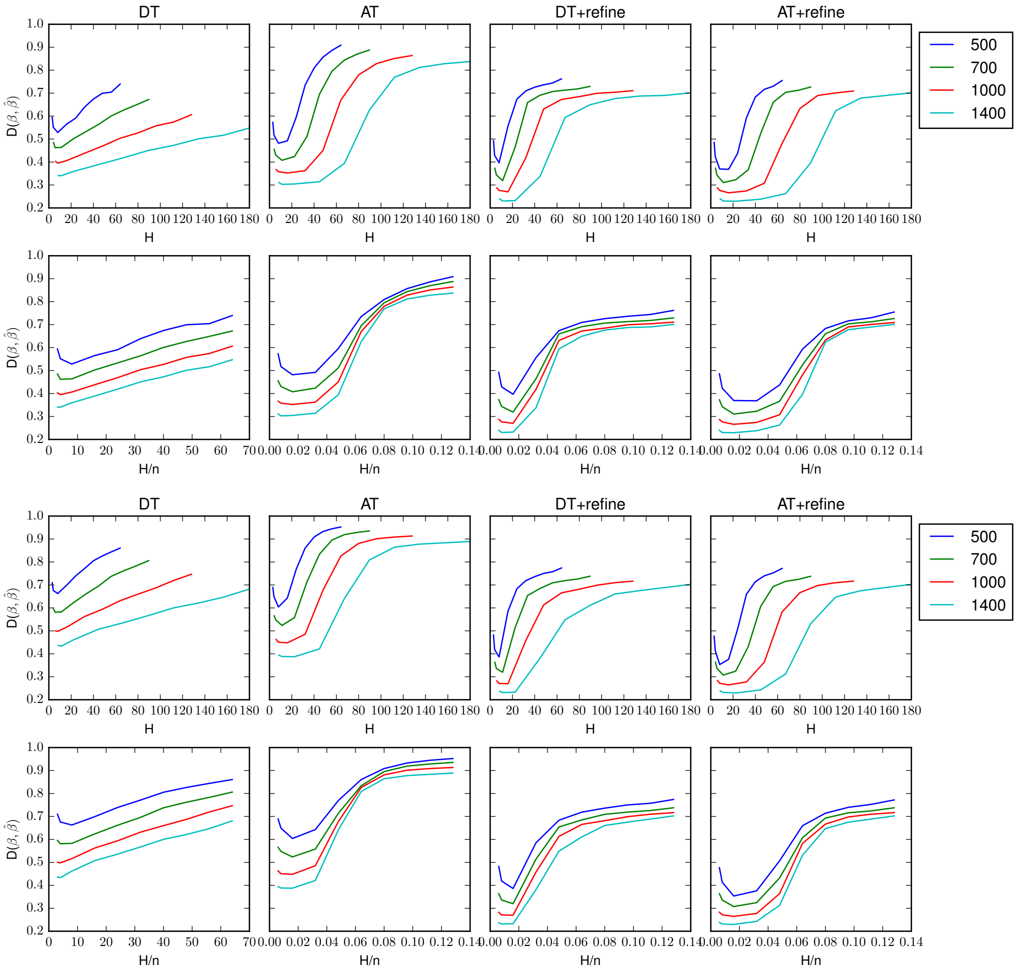


Figure 3.16: Model IV, top two rows: Σ_1, bottom two rows: Σ_2.

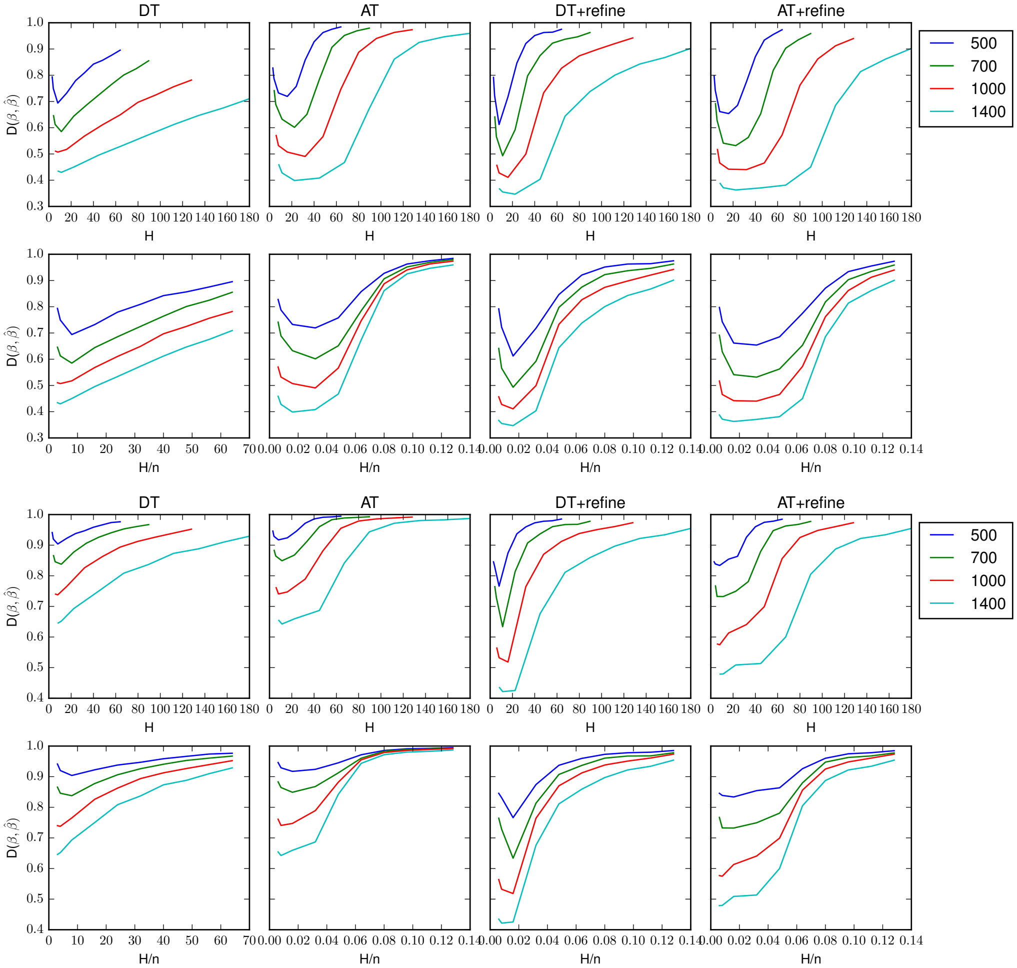


Figure 3.17: Model V, top two rows: Σ_1, bottom two rows: Σ_2.

3.5 Comparison of SSIR methods

In this section we compare SSIR-DT/AT with SSIR-SDP. For all methods, we run both with and without refinement. Simulation data are generated from Model I - IV with $\Sigma_X = \mathbf{I}_p$, and $(n,p) = (100,150),(200,300),(400,600)$. For SSIR-DT/AT, we first assume Σ_X is known; we use "*" to indicate the oracle version that true covariance is in use. We then plug in $\mathbf{X}^T\mathbf{X}/n$ and its pseudo-inverse into the algorithm.

Similarly as previous simulation, we only tune one parameter for each method. For DT, we set $\gamma_1 = (0.05, 0.1, ..., 1)\sqrt{\frac{D(1.5\log p + H/2)}{n}}$; for AT, we fix $\gamma_1 = 0.5\sqrt{\frac{D(1.5\log p + H/2)}{n}}$ and set $\gamma_2 = 0.5 \times 1.4^{(-10,...,10)}\frac{2D\log p}{n}$; for DT-ref, we fix $\gamma_1 = 0.5\sqrt{\frac{D(1.5\log p + H/2)}{n}}$ and set the tuning parameter in refinement $\lambda = 0.5 \times 1.4^{(-10,...,10)}\sqrt{\frac{D\log p}{n}}$; for AT-ref, we fix $\gamma_1 = 0.5\sqrt{\frac{D(1.5\log p + H/2)}{n}}$, $\gamma_2 = \frac{D\log p}{n}$, and set $\lambda = 0.5 \times 1.4^{(-10,...,10)}\sqrt{\frac{D\log p}{n}}$. For SSIR-SDP, we set $\rho = (0.1, ..., 1)\sqrt{\frac{\log p}{n}}$, and for SSIR-SDP-ref we fix $\rho = 0.5\sqrt{\frac{\log p}{n}}$ and set $\lambda = (0.1, ..., 2)\sqrt{\frac{\log p}{n}}$.

We get the minimal mean error from 200 independent runs of each method under each model configuration by tuning the one parameter. Results are shown in the following table. When $(n,p) = (100,150)$, SSIR-SDP performs better; the errors of SSIR-DT/AT are too large, so that after refinement the error is still much larger than SSIR-SDP. When (n,p) get larger, the errors of thresholded estimators shrinks rapidly, and after refinement, it is quite similar to SSIR-SDP. This shows that (i) thresholding-based methods require larger (n,p) to start converging than SSIR-SDP does; (ii) refinement can drastically improve the performance; and (iii) sample covariance and its pseudo-inverse are not good estimators of the covariance and precision matrix, so when p is large, the difference between using true covariance and estimated covariance becomes substantial.

Model	(n,p)	DT*	AT*	DT-ref*	AT-ref*	DT	AT	DT-ref	AT-ref	SDP	SDP-ref
I	(100,150)	0.607	0.662	0.327	0.328	0.677	0.482	0.382	0.353	0.172	0.174
	(200,300)	0.314	0.296	0.095	0.097	0.555	0.222	0.193	0.184	0.136	0.136
	(400,600)	0.175	0.144	0.095	0.091	0.295	0.064	0.077	0.072	0.080	0.080
II	(100,150)	0.537	0.564	0.184	0.189	0.533	0.303	0.284	0.242	0.081	0.085
	(200,300)	0.363	0.305	0.078	0.074	0.342	0.078	0.110	0.092	0.065	0.065
	(400,600)	0.209	0.152	0.041	0.037	0.209	0.025	0.045	0.037	0.033	0.033
III	(100,150)	0.732	0.797	0.656	0.661	0.903	0.956	0.785	0.785	0.375	0.362
	(200,300)	0.443	0.365	0.236	0.232	0.660	0.561	0.406	0.405	0.182	0.179
	(400,600)	0.158	0.185	0.141	0.137	0.464	0.408	0.446	0.429	0.155	0.119
IV	(100,150)	0.783	0.819	0.710	0.712	0.913	0.959	0.786	0.786	0.531	0.534
	(200,300)	0.585	0.594	0.517	0.516	0.735	0.685	0.619	0.619	0.393	0.393
	(400,600)	0.141	0.154	0.130	0.122	0.447	0.424	0.395	0.394	0.172	0.129

Table 3.1: Comparison between various SSIR methods.

We also want to mention that the result can be quite sensitive to tuning parameters. We can see that the error of DT/AT/DT-ref/AT-ref is sometimes much worse than their counterparts using the true covariance matrix. This is not surprising, but the error can be much smaller by tuning all available parameters (recall that we only tune one parameter). In practice, since thresholding-based methods run much faster than SSIR-SDP, it is feasible computationally.

In summary, SSIR-SDP performs consistently better, but it is more complex computationally than thresholding-based method. Thresholded estimator with refinement potentially can have similar performance as SSIR-SDP, but (n, p) cannot be too small and we need better estimators of covariance and precision matrices than sample covariance and its pseudo-inverse.

CHAPTER IV

Regression with Block Sparsity

4.1 Introduction

Linear regression is the most basic tool for studying how one random variable (response) is influenced by other random variables (predictors). Sometimes instead of random variables, one would be more interested in the relationship between more complicated random objects, like random vectors or random functions. Consider a linear model with multivariate response and grouped predictors, i.e.,

$$\mathbf{Y} = \mathbf{X}B + \mathbf{E} = \sum_{k=1}^{p} \mathbf{X}_{[k]} B_{[k]} + \mathbf{E} ,$$

where $\mathbf{Y}, \mathbf{E} \in \mathbb{R}^{n \times D_y}$, $\mathbf{X}_{[k]} \in \mathbb{R}^{n \times D_k}$ and $B_{[k]} = \mathbb{R}^{D_k \times D_y}$. When p is large, like ordinary regression, model selection becomes a important aspect of statistical modeling. It is desirable to identify those $X_{[k]}$'s that contribute to Y; also the ordinary estimator is not stable, and restricting to a small subset of predictors improves stability.

Given that group structure is known, it is preferable that $\hat{B}$ is sparse in the sense that the subset $\hat{S} = \{k : \hat{B}_{[k]} \neq \mathbf{0}_{D_k \times D_y}\}$ is small. We call this block sparsity. One can easily achieve block sparsity by adding a block norm penalty, that is:

$$\hat{B} = \operatorname*{arg\,min}_{B_{[k]} \in \mathbb{R}^{D_k \times D_y}} \frac{1}{2n} \left\| \mathbf{Y} - \sum_{k=1}^{p} \mathbf{X}_{[k]} B_{[k]} \right\|_{\mathrm{F}}^{2} + \lambda_n \sum_{k=1}^{p} \left\| B_{[k]} \right\|_{\mathrm{F}} , \tag{4.1}$$

$$\hat{S} = \{k : \hat{B}_{[k]} \neq \mathbf{0}_{D_k \times D_y}\} .$$

Obviously, this estimator is an extension of Lasso estimator: when $D_1 = ... = D_p = D_y = 1$, the penalty function reduces to ℓ_1 norm. Statistical property of Lasso estimator has been widely studied under different contexts. Knight and Fu (2000) provide asymptotic distribution of Lasso estimator; Greenshtein and Ritov (2004) prove the consistency of prediction under mild condition. To prove the consistency of model selection, i.e. $P(\hat{S} \approx S) \to 1$, there are essentially two techniques. One technique is to use some irrepresentable-type condition and prove exact sparsity recover. Zhao and Yu (2006) and Meinshausen and Bühlmann (2006) prove model selection consistency under fixed design and random design, respectively. Wainwright (2009a) provide a proof that is more non-asymptotic, although the proof still requires n and p to be large enough, without specifying how large they need to be. Another contribution of that paper is to provide necessary conditions of model s-election consistency, and compare the sample size required for consistency with an information-theoretic bound Wainwright (2009b). Another technique is to prove oracle inequality under some sparse eigenvalue or restricted eigenvalue conditions. This is closely related to the Danztig selector. The idea is to bound the coefficient estimation error in ℓ_2 norm or some other norm; if the ℓ_2 norm error is bounded, then there are not too many large coefficients being missed. This is called approximate sparsity recovery and is more realistic when there are a lot of small but non-zero coefficients. We refer to Candes and Tao (2007), Bickel et al. (2009), Zhou (2009) and reference therein for theoretical results and proof techniques of this type and van de Geer and Bühlmann (2009) for a discussion on the comparison between different variants of irrepresentable conditions and restricted eigenvalue conditions.

There have also been a lot of works on sparse regression that consider groups of coefficients. For example, if $D_y = 1$, but $D_k > 1, k \leq p$, (4.1) becomes group Lasso estimator. Asymptotic theory of group Lasso has been studied by Bach (2008), Nardi and Rinaldo (2008) which work only if n is large compared to p. Huang and Zhang (2010), Meier et al. (2009) also provide some statistical properties, but they have made some assumptions on the design that are suitable for their specific problems. If $D_y > 1$, but $D_k = 1, k \leq p$, then (4.1) becomes an estimator of multivariate (response) regression, where predictors share common support. Obozinski et al. (2011) use irrepresentable condition and Karim Lounici and Tsybakov (2011) us restricted eigenvalue condition to prove statistical consistency of this etimator. Obozinski et al. (2011) actually follows Wainwright (2009a) closely, and they suffer the same drawback that it is not specified how large (n, p, s) need to be for in their theorems. There are also works to unify all the theories, because the proof techniques for Lasso and group Lasso are similar, see for example Zhao et al. (2009), Negahban et al. (2012).

In this chapter, we prove model selection consistency of the block-penalized regression under random design in a non-asymptotic way. It is not hard to transform a non-asymptotic theory to an asymptotic one, and also not hard to adapt the proof to work for fixed-design setting. Note that Lasso and group Lasso are special cases of block Lasso, and our theoretical results are comparable to the theories for Lasso and group Lasso in the literature asymptotically. Our proof combines various proof techniques and is modular, so that it is easily understandable. In addition, we provide a lower bound on the penalty parameter, which is also purely non-asymptotic and obtained using a new proof technique.

4.2 Statistical consistency

4.2.1 Notation

Recall that for $\Sigma \in \mathbb{R}^{m\times m}$, $X \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{m\times D_y}$ where the m coordinates are divided into p groups $\{1, 2, ..., m\} = \uplus_{k=1}^{p} G_k$, $|G_k| = D_k$. We denote $\Sigma_{[jk]} = \Sigma_{G_j G_k}$, $X_{[k]} = X_{\cdot G_k}$ and $B_{[k]} = B_{G_k \cdot}$. Also in the bracket, we can use a set instead of a number to indicate a set of row groups or column groups.

Also we define the following norms that use Frobenius norm of sub-blocks as basic units. For example.

$$\|B\|_{1,F} = \sum_k \left\|B_{[k]}\right\|_{\mathrm{F}} \ ,$$

$$\|B\|_{\infty,F} = \max_k \left\|B_{[k]}\right\|_{\mathrm{F}} \ ,$$

$$\|\Sigma\|_{\infty\to\infty,F} = \max_k \max_{B:\|B\|_{\infty,F}\leq 1} \left\|[\Sigma B]_{[k]}\right\|_{\mathrm{F}} \ .$$

4.2.2 Exact sparsity recovery

Recall the linear model

$$\mathbf{Y} = \mathbf{X}B + \mathbf{E} = \sum_{k=1}^{p} \mathbf{X}_{[k]} B_{[k]} + \mathbf{E} \ .$$

Assume that the rows of $\mathbf{X}$ and $\mathbf{E}$ are i.i.d. samples from multivariate normal distributions that have 0 means and covariance matrices Σ and Σ_E, respectively. Let $S = \{k : B_{[k]} \neq \mathbf{0}\}$ be the block-wise sparse set of group indices that have non-zero coefficients. The total number of groups p can be large compared with n, but we assume S to be of size $s = |S|$, which is small.

Here we provide some sufficient conditions for the block-penalized estimator defined in (4.1) to exactly recovery the sparse subset, i.e. $P(\hat{S} = S) \to 1$. Such consistency in sparsity has been studied thoroughly for the classical Lasso regression, and the theoretical results provided below extends that. It has been widely known that for Lasso to achieve exact sparsity recovery, some irrepresentable condition is both sufficient and necessary. We define irrepresentable condition for our problem as follows.

Assumption IV.1. *We say Σ satisfies (group-wise) irrespresentable condition if there exists $\gamma > 0$, such that*

$$\max_{k \in S^c} \left\{ \left\| \sum_{k' \in S} [\Sigma_{[S^c S]} (\Sigma_{[SS]})^{-1}]_{[kk']} U_{[k']} \right\|_F \right\} \leq 1 - \gamma \tag{4.2}$$

for any $U_{[k']} \in \mathbb{R}^{D_{k'} \times D_y}$ such that $\|U_{[k']}\|_F \leq 1$.

Remark IV.2. *Note that $[\Sigma_{[S^c S]}(\Sigma_{[SS]})^{-1}]_{[kk']}$ is the the coefficients of $X_{[k']}$ in the regression of $X_{[k]}$ to $X_{[S^c]}$. Therefore, the assumption, like other irrepresentable conditions in the literature, is to assume that the "irrelevant variables" $X_[S^c]$ is not too correlated with $X_[S]$.*

In classical regression, i.e. $D_k = 1$ for all k and $D_y = 1$, the above assumption reduces to uniform irrepresentable condition for Lasso. We emphasize that here the assumption is made on the covariance matrix Σ, so it is under the random-design framework, as in Meinshausen and Bühlmann (2006), Wainwright (2009a). Zhao and Yu (2006) use similar condition under the fixed-design framework where the assumption is made on the Gram matrix $\Sigma^n = \mathbf{X}^T\mathbf{X}/n$ instead. Generally speaking, proofs under random-design framework involve more technicalities than those under the fixed-design framework.

Even for classical Lasso regression, there are many variants of the irrepresentable conditions and the uniform irrepresentable condition is just one of them. When $D = 1$, $U_{[k]}$'s are scalars, and the uniform irrepresentable condition is basically to require that the inequality holds for any $U_k \in [-1, 1]$. However, more commonly the inequality is only required to hold when $U_k = \mathbf{sign}(\beta_k)$, where β_k's are the *true regression coefficients. The assumption is then called the strong irrepresentable condition and is actually weaker than the uniform irrepresentable condition. Since here $U_{[k]}$'s are matrices, there is no easy definition of a sign function, and hence it is natural to adapt the uniform condition and the inequality is required to hold for $\|U_{[k]}\|_F \leq 1$. Some other options can result in weaker assumptions; e.g. i) $\|U_{[k]}\|_F = 1$, ii) $\|U_{[k]}\|_F \leq 1$ and $\langle U_{[k]}, B_{[k]}\rangle \geq 0$.*

The condition can actually be satisfied. For example, if $A \in \mathbb{R}^{p\times p}$ satisfies the uniform irrepresentable condition for classic Lasso, then for any positive definite matrix $B \in \mathbb{R}^{D\times D}$, $\Sigma = B \otimes A$ satisfies the group-wise irrepresentable condition. The most favorable case is identity matrix $\Sigma = c\mathbf{I}_{pD}$, where the assumption is satisfied with $\gamma = 1$.

We use the well-known proof technique called the primal-dual witness (PDW) construction, described in the following proposition

Proposition IV.3. *For a subset $S \subseteq \{1, 2, ..., p\}$, consider $\tilde{B}$ such that*

$$\tilde{B}_{[S]} = \underset{B\in\mathbb{R}^{D_k\times D_y}, k\in S}{\arg\min} \frac{1}{2n}\left\|\mathbf{Y} - \sum_{k\in S}\mathbf{X}_{[k]}B_{[k]}\right\|_{\mathrm{F}}^2 + \lambda_n \sum_{k\in S}\left\|B_{[k]}\right\|_{\mathrm{F}}, \tag{4.3}$$

and $\tilde{B}_{[k']} = \mathbf{0}_{D_{k'}\times D_y}$ for $k' \in S^c$.

1. *Denote the residual of the restricted problem by $\tilde{\mathbf{R}} = \mathbf{Y} - \mathbf{X}_{[S]}\tilde{B}_{[S]}$. If $\|\tilde{\mathbf{R}}^T\mathbf{X}_{[k']}\|_F < \lambda_n$ for all $k' \in S^c$, then $\tilde{B}$ is also a solution of the non-restricted problem (4.1).*

2. *Moreover, if $\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}$ is not singular, then $\tilde{B}$ is the unique solution.*

Proof. Let the objective function of the restricted and non-restricted problem be

$$h_1(B_{[S]}) = \frac{1}{2n}\left\|\mathbf{Y} - \sum_{k\in S}\mathbf{X}_{[k]}B_{[k]}\right\|_{\mathrm{F}}^2 + \lambda_n \sum_{k\in S}\left\|B_{[k]}\right\|_{\mathrm{F}}$$

and

$$h_2(B_{[S]}, B_{[S^c]}) = \frac{1}{2n}\left\|\mathbf{Y} - \sum_{k\in S}\mathbf{X}_{[k]}B_{[k]} - \sum_{k'\in S^c}\mathbf{X}_{[k']}B_{[k']}\right\|_{\mathrm{F}}^2 + \lambda_n \sum_{k\in S}\left\|B_{[k]}\right\|_{\mathrm{F}} + \lambda_n \sum_{k'\in S}\left\|B_{[k']}\right\|_{\mathrm{F}} \ ,$$

respectively. For any optimal solution $\tilde{B}_{[S]}$ of $\arg\min h_1(B_{[S]})$, by first-order optimality condition of sub-differentiable convex function, we know that the sub-gradient of h_1 at $\tilde{B}_{[S]}$, denoted by $\partial h_1|_{\tilde{B}_{[S]}}$, includes zero. To prove that the zero-padded matrix $\tilde{B}$ is a minimal solution of $\arg\min h_2$, by first-order optimality condition again, we only need to show that $\mathbf{0} \in \partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}$. Since $h_2(B_{[S]}, \mathbf{0}) = h_1(B_{[S]})$ for any $B_{[S]} \in \mathbb{R}^{D_S\times D_y}$, we have

$$\left[\partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}\right]_{[S]} = \partial h_1|_{\tilde{B}_{[S]}} \ .$$

We need to emphasize that sub-gradient is a set of matrices, not a single matrix; the equality above means that the two sets are the same. We already know that $\mathbf{0} \in \partial h_1|_{\tilde{B}_{[S]}}$. To show that $\mathbf{0} \in \partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}$, we only need to show that

$$\mathbf{0} \in \left[\partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}\right]_{[S^c]} \ ,$$

or equivalently

$$\mathbf{0} \in \left[\partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}\right]_{[k']} \quad \forall k' \in S^c \ .$$

Note that

$$\left[\partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}\right]_{[k']} = \left\{\frac{1}{n}\mathbf{X}_{[k']}^T\tilde{\mathbf{R}} + \lambda_n U \,:\, U \in \mathbb{R}^{D_{k'}\times D_y}, \|U\|_F \le 1\right\} \ .$$

Since $\left\|\tilde{\mathbf{R}}^T\mathbf{X}_{[k']}/n\right\|_{\mathrm{F}} < \lambda_n$, we have $\mathbf{0} \in \left[\partial h_2|_{(\tilde{B}_{[S]},\mathbf{0})}\right]_{[k']}$, which finish the proof.

To prove uniqueness, first notice that even if the minimizer of h_2 is not unique, the prediction $\mathbf{X}\tilde{B}$ is unique. This is because if we have two distinct minimizers $\tilde{B}^{(1)}$ and $\tilde{B}^{(2)}$, then letting $\tilde{B}^{(3)} = (\tilde{B}^{(1)} + \tilde{B}^{(2)})/2$, we have

$$2h_2(\tilde{B}^{(3)}) \le h_2(\tilde{B}^{(2)}) + h_2(\tilde{B}^{(1)})$$

due to convexity. Since $h_2(\tilde{B}^{(1)})$ and $h_2(\tilde{B}^{(2)})$ are already minimal, the equality must hold; and since h_2 is the sum of two convex functions, the equality has to hold for each convex function. Therefore,

$$\left\|\mathbf{Y} - \mathbf{X}\tilde{B}^{(1)}\right\|_F^2 + \left\|\mathbf{Y} - \mathbf{X}\tilde{B}^{(2)}\right\|_F^2 = 2\left\|\mathbf{Y} - \mathbf{X}\tilde{B}^{(3)}\right\|_F^2 .$$

This is only possible if $\mathbf{X}\tilde{B}^{(1)} = \mathbf{X}\tilde{B}^{(2)}$ (because $\|\mathbf{Y} - \mathbf{Z}\|_F$, as a function of $\mathbf{Z}$, is strictly convex). Thus the residual $\tilde{\mathbf{R}}$ is also unique. This means that the condition $\frac{1}{n}\left\|\tilde{\mathbf{R}}^T\mathbf{X}_{[k']}\right\|_F < \lambda_n$ holds for any minimal solution $\tilde{B}$ of h_2, which implies that $\tilde{B}_{[k']} = \mathbf{0}$ for any minimal solution. In conclusion any minimizer of h_2 is supported on S, so they are also minimizers of h_1, padded with zeros.

If further $\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}$ is not singular, then $h_1(B_{[S]})$ is the sum of a strictly convex function (the quadratic loss) and a convex function (the penalty), so the optimal solution of h_1, i.e., $\tilde{B}_{[S]}$, is unique. Thus, the minimizer of h_2 is also unique. □

With the proposition, proving sparsity recovery is equivalent to proving the following two events

$$\tilde{B}_{[k]} \neq \mathbf{0} \text{ for } k \in S \ , \tag{4.4}$$

$$\|\tilde{\mathbf{R}}^T\mathbf{X}_{[k]}\|_F < \lambda_n \text{ for all } k \in S^c \ . \tag{4.5}$$

The first event means no type I error or no false negatives, i.e. if a group of predictors has non-zero coefficients, it is included in $\hat{S}$; the second event means no type II error or no false positives, i.e. all groups of predictors that have zero coefficients are excluded from $\hat{S}$. If both events occur, then the sparse model is exactly recovered.

Thus, our main theorem is divided into two parts accordingly. On one hand, we need the penalty parameter λ_n to be sufficiently large compared with the noise, so that (4.5) holds with large probability; on the other hand, since the penalty also causes bias, we need λ_n to be not too large compared with the signal level, so that (4.4) holds with large probability.

The following theorem provides sufficient conditions for exact sparsity recovery.

Theorem IV.4. *Let* $D_{\max} = \max_k\{D_k\} \vee D_y$*, and* $D_S = \sum_{k\in S} D_k$*,* $\beta_{\min} = \min\{\|B_{[k]}\|_F : k \in S\}$*. We make the following assumptions:*

1. **Assumption IV.1** *with constant* γ*;*

2. **Regularity condition on the distribution of $\mathbf{X}, \mathbf{E}$:** $\Sigma_{[SS]}$ *is non-singular, and after proper scaling,* $\sigma_{\max}(\Sigma_{[kk]}) \leq 1$ *for any* $k \in S$*,* $\sigma_S^2 = \sigma_{\min}(\Sigma_{[SS]}) > 0$*,* $\theta_\infty = \left\|\Sigma_{[SS]}^{-1}\right\|_{\infty\to\infty,F}$*, and* $\sigma_E^2 = \sigma_{\max}(\Sigma_E)$*;*

3. **Condition on the problem size parameters:** *there exist absolute constant* $\delta, t, C_2 > 0$*,* $0 < C_1 < 1$*, such that*

$$\sqrt{n} \geq \frac{\sqrt{s}(\sqrt{D_{\max}} + \sqrt{2\log(p-s) - 2\log\delta})}{(\gamma/2)C_2\sigma_{[S]}} \,, \tag{4.6}$$

$$\frac{\sqrt{D_S \vee D_{\max}}}{\sqrt{n}} + t < C_1 \,, \tag{4.7}$$

$$\frac{\sqrt{D_S \vee D_{\max}}}{\sqrt{n}} + t < \frac{C_2}{\sqrt{s}} \,. \tag{4.8}$$

Under these assumptions, we have the following sparsity recovery properties:

1. *By choosing tuning parameter as*

$$\lambda_n = \frac{(1+C_1)\sigma_E}{\gamma/2} \frac{D_{\max} + \sqrt{2\log(p-s) - 2\log\delta}}{\sqrt{n}} \,, \tag{4.9}$$

we have

$$P\{\hat{S} \subseteq S\} \geq 1 - 4\exp(-t^2 n/2) - 2\delta \,.$$

2. Moreover, if

$$\beta_{\min} \geq \frac{\sigma_E}{\sigma_S(1-C_1)} \frac{(D_{\max} + \sqrt{2\log s - 2\log \delta})}{\sqrt{n}} + \left(\theta_\infty + \frac{(2-C_1)C_2}{(1-C_1)^2}\right) \lambda_n \ , \tag{4.10}$$

then the restricted problem (4.3) satisfies

$$P\left(\hat{S} \supseteq S\right) \geq 1 - 4\exp(-t^2 n/2) - 3\delta \ .$$

If $p > 2s$*, then (4.10) can be replaced by*

$$\beta_{\min} \geq \left(\frac{\gamma/2}{\sigma_S(1-C_1^2)} + \theta_\infty + \frac{(2-C_1)C_2}{(1-C_1)^2}\right) \lambda_n \ . \tag{4.10'}$$

3. With (4.9) and (4.10), we have:

$$P\left(\hat{S} = S\right) \geq 1 - 4\exp(-t^2 n/2) - 3\delta \ .$$

Proof. We divide the proof into 3 parts. For notational simplicity, we assume $D_1 = ... = D_p = D_Y = D$. The proof for unequal group sizes is not very different.

- **Bounding the deviation of $\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n$ from $\Sigma_{[SS]}$:**

Let $\mathbf{Z} = \mathbf{X}_{[S]} \Sigma_{[SS]}^{-1/2}$. Then $\mathbf{Z}$ is an n by sD independent Gaussian ensemble, so

$$\begin{aligned} & \left\|(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} - \Sigma_{[SS]}^{-1}\right\|_{\text{op}} \\ =& \left\|\Sigma_{[SS]}^{-1/2}\left((\mathbf{Z}^T\mathbf{Z}/n)^{-1} - I_{sD}\right)\Sigma_{[SS]}^{-1/2}\right\|_{\text{op}} \\ \leq& \left\|\Sigma_{[SS]}^{-1/2}\right\|_{\text{op}}^2 \left\|(\mathbf{Z}^T\mathbf{Z}/n)^{-1} - I_{sD}\right\|_{\text{op}} \\ =& \frac{1}{\sigma_S^2}\left|\frac{1}{\sigma_{\min}(\mathbf{Z}^T\mathbf{Z}/n)} - 1\right| \ , \end{aligned} \tag{4.11}$$

$$\left\|(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1}\right\|_{\text{op}} \leq \frac{1}{\sigma_S^2 \sigma_{\min}(\mathbf{Z}^T\mathbf{Z}/n)} \ . \tag{4.12}$$

By **Lemma II.14**, for any $t > 0$, with probability at most $2\exp(-t^2 n/2)$, the event

$$\mathcal{A} = \left\{ \left| \sqrt{\sigma_{\min}(\mathbf{Z}^T\mathbf{Z}/n)} - 1 \right| > \frac{\sqrt{sD}}{\sqrt{n}} + t \right\} ,$$

Holds. By our assumption, we can find $0 < C_1 < 1$ and $C_2 > 0$, such that

$$\frac{\sqrt{sD}}{\sqrt{n}} + t < C_1 \wedge \frac{C_2}{\sqrt{s}} .$$

Note that if $|u-1| \leq c < 1$, then

$$\left| \frac{1}{u^2} - 1 \right| = |1-u| \left| \frac{1}{u} + \frac{1}{u^2} \right| \leq \frac{(2-c)}{(1-c)^2} |1-u| .$$

Using this inequality, we have that event $\mathcal{A}^c$ implies

$$\sigma_{\min}(\mathbf{Z}^T\mathbf{Z}/n) \geq (1-C_1)^2 , \tag{4.13}$$

$$|1/\sigma_{\min}(\mathbf{Z}^T\mathbf{Z}/n) - 1| \leq \frac{2-C_1}{(1-C_1)^2} \frac{C_2}{\sqrt{s}} . \tag{4.14}$$

Combining (4.11) and (4.14), and combining (4.12) and (4.13), we have that $\mathcal{A}^c$ implies

$$\left\| (\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} \right\|_{\text{op}} \leq \frac{1}{\sigma_S^2 (1-C_1)^2} , \tag{4.15}$$

$$\left\| (\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} - \Sigma_{[SS]}^{-1} \right\|_{\text{op}} \leq \frac{(2-C_1)C_2}{\sigma_S^2 (1-C_1)^2} \frac{1}{\sqrt{s}} . \tag{4.16}$$

Later we will frequently condition on $\mathbf{X}$ and exclude $\mathcal{A}$. Inequality (4.15) is used in both parts of the theorem; inequality (4.16) is used in bounding a specific term in the second part of theorem.

- **Part one of the theorem:**

To prove the first part of the theorem, we use the irrepresentable condition.

Since $\tilde{B}_{[S]}$ minimizes (4.3), the sub-gradient of (4.3) at $\tilde{B}_{[S]}$ has to include $\mathbf{0}$. Thus we can find $\check{B}_{[k]} \in \mathbb{R}^{D_k \times D_y}$ for $k \in S$ satisfying

$$\begin{cases} \check{B}_{[k]} = \tilde{B}_{[k]} / \left\| \tilde{B}_{[k]} \right\|_{\mathrm{F}} & \tilde{B}_{[k]} \neq \mathbf{0} \\ \left\| \check{B}_{[k]} \right\|_{\mathrm{F}} \leq 1 & \tilde{B}_{[k]} = \mathbf{0} \end{cases}$$

such that

$$\frac{\mathbf{X}_{[k]}^T \tilde{\mathbf{R}}}{n} = \lambda_n \check{B}_{[k]} \ ,$$

where the residual is $\tilde{\mathbf{R}} = \mathbf{Y} - \mathbf{X}_{[S]} \tilde{B}_{[S]}$.

To utilize the random design irrepresentable condition, write the population regression model of $\mathbf{X}_{[k']}$, $k' \in S^c$ on $\mathbf{X}_{[S]}$ as

$$\mathbf{X}_{[k']} = \mathbf{X}_{[S]} (\Sigma_{[SS]})^{-1} \Sigma_{[Sk']} + \mathbf{V}_{[k']} \ . \tag{4.17}$$

With Gaussian assumption, $\mathbf{V}_{[k']}$'s are independent of $\mathbf{X}_{[S]}$. Thus

$$\begin{aligned} \mathbf{X}_{[k']}^T \tilde{\mathbf{R}} &= \Sigma_{[k'S]} \Sigma_{[SS]}^{-1} \mathbf{X}_{[S]}^T \tilde{\mathbf{R}} + \mathbf{V}_{[k']}^T \tilde{\mathbf{R}} \\ &= n \lambda_n \Sigma_{[k'S]} \Sigma_{[SS]}^{-1} \check{B}_{[S]} + \mathbf{V}_{[k']}^T \tilde{\mathbf{R}} \ . \end{aligned}$$

Therefore,

$$\left\| \frac{\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}}{n} \right\|_{\mathrm{F}} \leq \lambda_n \left\| \Sigma_{[k'S]} \Sigma_{[SS]}^{-1} \check{B}_{[S]} \right\|_{\mathrm{F}} + \left\| \frac{\mathbf{V}_{[k']}^T \tilde{\mathbf{R}}}{n} \right\|_{\mathrm{F}} \ .$$

By **Proposition IV.3**, we only need to prove that, with high probability,

$$\max_{k' \in S^c} \left\| \frac{\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}}{n} \right\|_{\mathrm{F}} \leq \lambda_n \ .$$

Using irrepresentable condition, it is then enough to prove that, with high probability,

$$\max_{k' \in S^c} \left\| \frac{\mathbf{V}_{[k']}^T \tilde{\mathbf{R}}}{n} \right\|_{\mathrm{F}} < \gamma \lambda_n \ .$$

Thus, we need to make λ_n large enough so that it bounds the quantity on the left with high probability. Note that

$$\tilde{\mathbf{R}} = \mathbf{Y} - \mathbf{X}_{[S]}\tilde{B}_{[S]} = \mathbf{X}_{[S]}(B_{[S]} - \tilde{B}_{[S]}) + \mathbf{E}$$

Let $P_{[S]} = \mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}\mathbf{X}_{[S]}^T$, the projection matrix onto the columns of $\mathbf{X}_{[S]}$; and let $P_{[S]}^{\perp} = \mathbf{I}_n - P_{[S]}$. Then

$$\begin{aligned}\tilde{\mathbf{R}} &= P_{[S]}\tilde{\mathbf{R}} + P_{[S]}^{\perp}\tilde{\mathbf{R}} = \mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}\mathbf{X}_{[S]}\tilde{\mathbf{R}} + P_{[S]}^{\perp}\mathbf{E} \\ &= n\lambda_n\mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}\check{B}_{[S]} + P_{[S]}^{\perp}\mathbf{E}\ .\end{aligned}$$

Denote $H = \sqrt{n}\mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}\check{B}_{[S]}$. Then

$$\frac{\mathbf{V}_{[k']}^T\tilde{\mathbf{R}}}{n} = \frac{\mathbf{V}_{[k']}^T H}{\sqrt{n}}\lambda_n + \frac{\mathbf{V}_{[k']}^T P_{[S]}^{\perp}\mathbf{E}}{n}\ .$$

Note that H is decided by $\mathbf{X}_{[S]}, \mathbf{E}$. Thus it is independent of $\mathbf{V}_{[k']}$ for any $k' \in S^c$. Although H does not have i.i.d. rows, by conditioning on $\mathbf{X}_{[S]}, \mathbf{E}$, the concentration of $\mathbf{V}_{[k']}^T H$ is described by (4.20) in **Lemma IV.8**

$$P\left(\left\|\mathbf{V}_{[k']}^T H\right\|_{\mathrm{F}} > (\sqrt{D} + \sqrt{2t_2})\left\|H\right\|_{\mathrm{F}}\ \middle|\ \mathbf{X}, \mathbf{E}\right) \le \exp(-t_2)$$

To bound $\|H\|_{\mathrm{F}}$, we have

$$H^T H = \check{B}_{[S]}^T(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\ ,$$

hence

$$\|H\|_{\mathrm{F}}^2 \le \left\|\check{B}_{[S]}\right\|_{\mathrm{F}}^2 \left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\right\|_{\mathrm{op}}\ .$$

We know that $\left\|\check{B}_{[S]}\right\|_{\mathrm{F}}^2 = \sum_{k\in S}\left\|\check{B}_{[k]}\right\|_{\mathrm{F}}^2 \le s$. Also, from the definition of $\mathcal{A}$, we have

$$P\left(\left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\right\|_{\mathrm{op}} > \frac{1}{(1-C_1)^2\sigma_S^2}\right) < 2\exp(-t^2 n/2)\ .$$

Therefore, in summary, we have

$$P\left(\forall k' \in S^c, \left\|\mathbf{V}_{[k']}^T H\right\|_{\mathrm{F}} > \frac{\sqrt{s}(\sqrt{D}+\sqrt{2t_2})}{(1-C_1)\sigma_S}\right) \\ \leq 2\exp(-t^2 n/2) + (p-s)\exp(-t_2)\ .$$

On the other hand, $P^\perp$ is decided by $\mathbf{X}_{[S]}$, so $P^\perp$, $\mathbf{E}$ and $\mathbf{V}_{[k']}$ are mutually independent. Also note that $\sigma_{\max}(\mathbf{var}(V_{[k']})) \leq \sigma_{\max}(\mathbf{var}(X_{[k']})) = 1$. We can use (4.19) in **Lemma IV.8**

$$P\left(\sqrt{n}\left\|\mathbf{V}_{[k']}^T P^\perp \mathbf{E}/n\right\|_{\mathrm{F}} > (D+\sqrt{2t_2})\left\|P^\perp\mathbf{E}/\sqrt{n}\right\|_{\mathrm{op}} \middle| \mathbf{X}, \mathbf{E}\right) \leq \exp(-t_2)\ .$$

By **Lemma II.14**

$$P\left(\left\|\mathbf{E}/\sqrt{n}\right\|_{\mathrm{op}} > (1+\sqrt{D/n}+t)\sigma_E\right) \leq 2\exp(-t^2 n/2)\ .$$

Since $\left\|P^\perp\mathbf{E}/\sqrt{n}\right\|_{\mathrm{op}} \leq \|\mathbf{E}/\sqrt{n}\|_{\mathrm{op}}$,

$$P\left(\left\|P^\perp\mathbf{E}/\sqrt{n}\right\|_{\mathrm{op}} > (1+C_1)\sigma_E\right) \leq 2\exp(-t^2 n/2)\ .$$

Therefore,

$$P\left(\forall k' \in S^c, \sqrt{n}\left\|\mathbf{V}_{[k']}^T P^\perp\mathbf{E}/n\right\|_{\mathrm{F}} > (1+C_1)\sigma_E(D+\sqrt{2t_2})\right) \\ \leq 2\exp(-t^2 n/2) + (p-s)\exp(-t_2)\ .$$

In summary, letting $t_2 = \log((p-s)/\delta)$ and assuming

$$\sqrt{n} \geq \frac{\sqrt{s}(\sqrt{D}+\sqrt{2\log(p-s)-2\log\delta})}{(\gamma/2)(1-C_1)\sigma_S}\ .$$

we can select

$$\lambda_n = \frac{(1+C_1)\sigma_E(D+\sqrt{2\log(p-s)-2\log\delta})}{(\gamma/2)\sqrt{n}}\ ,$$

such that

$$\max_{k'\in S^c} \frac{\mathbf{V}_{[k']}^T\tilde{\mathbf{R}}}{n} < \gamma\lambda_n$$

with probability at least

$$1 - 4\exp(-t^2 n/2) - 2\delta \ .$$

- **Part two of the theorem:**

Recall that $\mathbf{X}_{[k]}^T\tilde{\mathbf{R}} = n\lambda_n \check{B}_{[k]}$, and $\tilde{\mathbf{R}} = \mathbf{Y} - \mathbf{X}_{[S]}\tilde{B}_{[S]}$. Since $\Sigma_{[SS]}$ is non-singular, with probability one, $\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}$ is invertible. We then have

$$\tilde{B}_{[S]} = (\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}(\mathbf{X}_{[S]}^T\mathbf{Y} - n\lambda_n\check{B}_{[S]}) \ ,$$

and hence

(4.18)

$$\begin{aligned} &\tilde{B}_{[S]} - B_{[S]} \\ =&(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}(\mathbf{X}_{[S]}^T\mathbf{E}/n - \lambda_n\check{B}_{[S]}) \\ =&(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}(\mathbf{X}_{[S]}^T\mathbf{E}) - \lambda_n\Sigma_{[SS]}^{-1}\check{B}_{[S]} + \lambda_n\left((\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1} - \Sigma_{[SS]}^{-1}\right)\check{B}_{[S]} \ . \end{aligned}$$

The goal is to prove that $\tilde{B}_{[k]} \neq 0$ for all $k \in S$. It is sufficient to prove that $\left\|\tilde{B}_{[k]} - B_{[k]}\right\|_{\mathrm{F}} < \left\|B_{[k]}\right\|_{\mathrm{F}}$ for all $k \in S$, or to prove that the max block norm of $\tilde{B}_{[S]} - B_{[S]}$ is smaller than $\beta_{\min} = \min_{k\in S}\left\|B_{[k]}\right\|_{\mathrm{F}}$. The strategy is to bound the max block norm of these three parts separately.

To bound the first term, denote $J_{[k]} = \begin{pmatrix} \mathbf{0}_{D\times D} \\ \vdots \\ \mathbf{I}_D \\ \vdots \\ \mathbf{0}_{D\times D} \end{pmatrix}$, $H_{[k]} = \mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}J_{[k]}$.

Then the k-th block in the first term is

$$\left[(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}(\mathbf{X}_{[S]}^T\mathbf{E})\right]_{[k]} = \frac{1}{n}H_{[k]}^T\mathbf{E} \ .$$

Note that

$$H_{[k]}^T H_{[k]}/n = J_{[k]}^T (\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} J_{[k]} \ .$$

Thus for any $k \in \{1, ..., s\}$, we have $\left\| H_{[k]}^T H_{[k]}/n \right\|_{\text{op}} \leq \left\| (\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} \right\|_{\text{op}}$. Thus, we can condition on $\mathbf{X}$, exclude $\mathcal{A}$, and use **Lemma IV.8** to get

$$P\left(n \left\| H_{[k]}^T \mathbf{E}/n \right\|_{\text{F}}^2 > \frac{\sigma_E^2}{\sigma_S^2 (1-C_1)^2} (D + \sqrt{2t_3})^2 \right) \leq 2\exp(-nt^2/2) + \exp(-t_3) \ .$$

When bounding $\max_{k \in S} n \left\| H_{[k]}^T \mathbf{E}/n \right\|_{\text{F}}^2$, we can use Bonferroni bound and multiply the probability above by s. However, the $2\exp(-t^2 n/2)$ term is the probability bound of $\mathcal{A}$, and we do not need to count the probability repeatedly when taking sum of the probabilities of sub-events. Therefore,

$$\begin{aligned} &P\left(\left\| (\mathbf{X}_{[S]}^T \mathbf{X}_{[S]})^{-1} \mathbf{X}_{[S]} \mathbf{E} \right\|_{\infty,F} > \frac{\sigma_E}{\sigma_S(1-c_3)} \frac{(D + \sqrt{2t_3})}{\sqrt{n}} \right) \\ &\leq 2\exp(-nt^2/2) + s\exp(-t_3) \ . \end{aligned}$$

The second term is addressed by our assumption

$$\left\| \lambda_n \Sigma_{[SS]}^{-1} \check{B}_{[S]} \right\|_{\infty,F} \leq \left\| \Sigma_{[SS]}^{-1} \right\|_{\infty \to \infty, F} \lambda_n = \theta_\infty \lambda_n$$

For the third term, since we restrict to $\mathcal{A}^c$, from (4.16)

$$\begin{aligned} &\left\| \lambda_n \left((\mathbf{X}_{[S]}^T \mathbf{X}/n) - \Sigma_{[SS]}^{-1} \right) \check{B}_{[S]} \right\|_{\infty,F} \\ \leq &\left\| \lambda_n \left((\mathbf{X}_{[S]}^T \mathbf{X}/n) - \Sigma_{[SS]}^{-1} \right) \check{B}_{[S]} \right\|_{\text{F}} \\ \leq &\lambda_n \left\| (\mathbf{X}_{[S]}^T \mathbf{X}/n) - \Sigma_{[SS]}^{-1} \right\|_{\text{op}} \left\| \check{B}_{[S]} \right\|_{\text{F}} \\ \leq &\frac{(2-C_1)C_2}{(1-C_1)^2} \lambda_n \end{aligned}$$

In summary, letting $t_3 = \log(s/\delta)$, for any λ_n the probability of

$$\begin{aligned} \left\| \tilde{B}_{[S]} - B_{[S]} \right\|_{\infty,F} \geq &\frac{\sigma_E}{\sigma_S(1-C_1)} \frac{(D + \sqrt{2\log s - 2\log \delta})}{\sqrt{n}} \\ &+ \left(\theta_\infty + \frac{(2-C_1)C_2}{(1-C_1)^2} \right) \lambda_n \end{aligned}$$

is at most

$$2\exp(-nt^2/2) + \delta\ .$$

To prove part 3, we just need to use union bound and take sum of the probabilities in the first two parts. Note that $4\exp(-t^2n/2)$ in part 1 includes the probability of $\mathcal{A}$, which is where the $2\exp(-t^2n/2)$ in part 2 is from, so we do not count that probability twice. □

The above theorem is non-asymptotic and all the constants can be pre-specified. Moreover the high probability bound holds for any combination of (n, p, s, D) as long as it satisfies (4.6)-(4.8).

Remark IV.5 (Remarks on the assumptions)**.** *In* **Theorem IV.4** *we made two extra sets of assumptions besides the irrepresentable condition.*

The first set of assumptions are regularity conditions on the distribution of $\mathbf{X}$ *and* $\mathbf{E}$*. First, we assume that* X *is properly scaled so that each group* $X_{[k]}$ *has a covariance matrix whose operator norm is 1. With such assumption,* $B_{[k]}$*'s are comparable across different k's; without such assumption, it is not reasonable to use an equal* λ_n *to penalize all groups. It might be more natural to assume that each individual coordinate is of unit variance because normalization is done within single columns in practice. We did not explore that direction. Second, we assume that the true predictors* $X_{[S]}$ *are not singular; otherwise, the coefficients* $B_{[S]}$ *are not unique. Once assuming non-singularity, the operator norm of* $\Sigma_{[SS]}^{-1}$ *w.r.t. Frobenius norm (*σ_S^{-2}*) and max block norm (*θ_∞*) are well defined and present in the final non-asymptotic bound.*

The second set of assumptions specify the problem size regime to which the theory applies. We have three requirements on (n, p, s, D)*. The first requirement (4.6) is* *used when applying the irrepresentable condition; the second requirement (4.7) is used to bound $\sigma_{\min}(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)$ away from 0; the third requirement (4.8) is used when bounding the max block norm of $\tilde{B}_{[S]} - B_{[S]}$; None of these requirements are asymptotic.*

The non-asymptotic results can be easily modified to an asymptotic one. The following corollary gives an example.

Corollary IV.6. *If*

$$\liminf \frac{n}{s^2 D + s \log p + \frac{D^2 + \log p}{\beta_{\min}^2}} > \bar{C}$$

for some large enough constant $\bar{C}$, then there exists λ_n so that $\hat{S}$ satisfies $P(\hat{S} = S) \to 1$.

If $\beta_{\min} \asymp 1/\sqrt{s}$, then it is sufficient to have

$$\liminf \frac{n}{s^2 D + s D^2 + s \log p} > \bar{C}$$

Note that the third sample size assumption (4.8) may not be necessary. In the proof, we need this requirement to make the deviation $\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n - \Sigma_{[SS]}$ smaller, so that the Frobenius norm of $((\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} - \Sigma_{[SS]}^{-1})\check{B}_{[S]}$ is bounded by constant ($\check{B}_{[S]}$ is defined in the proof of Theorem IV.4); however, what we actually need is that the max block norm of it is be bounded. These two norms can have a ratio of as large as $\sqrt{s}$, which is why there is an extra $\sqrt{s}$ on the r.h.s. of (4.8). If we can get rid of this assumption, then the asymptotic sample size requirement can be simplified to $n \succ sD^2 + s \log p$. We have a dedicated simulation to show that very likely we can avoid the extra term (see Appendix 4.5), and numerical study also validates this conjecture.

4.2.3 Lower bound of λ_n

Two things are revealed in **Theorem IV.4**: the choice of λ_n and the corresponding $\beta_{\min}$ condition. These are also two important aspects in sparsity recovery of classical Lasso regression.

A smaller λ_n is preferred because then the bias is smaller and consequently the $\beta_{\min}$ can be smaller. According to Theorem IV.4, it is sufficient to choose $\lambda_n \succeq \frac{D_{\max}+\sqrt{\log p}}{\sqrt{n}}$. In the following theorem, we prove in a specific case that the asymptotic rate of this λ_n is necessary; smaller λ_n might cause non-diminishing probability of false inclusion.

Theorem IV.7 (Lower bound on λ_n)**.** *Assume that $D_1 = D_2 = ... = D_p = D_y = D$, and $\mathbf{cov}(X_{[S]}) = I_{sD\times sD}$, and $\mathbf{cov}(E) = \sigma^2 I_{D\times D}$. If $\frac{n}{sD} \wedge (\frac{n}{D} - s) \geq 1 + c_1 > 1$, and if $p - s$ is not too small (e.g. larger than 7). Then exist constants C_3, C_4, C_5 such that the following holds: if*

$$\lambda_n \leq \frac{C_3(D + \sqrt{C_4 \log(p-s)})}{\sqrt{n}},$$

then

$$P(\hat{S} \not\subseteq S) \geq C_5 > 0.$$

Proof. Note that the residual of the oracle problem (4.3), $\tilde{\mathbf{R}}$, only depends on $\mathbf{X}_{[k]}$, $k \in S$ and $\mathbf{E}$; thus it is independent of $\mathbf{X}_{[k']}$, $k' \in S^c$. Thus, for any constant $c > 0$,

$$\begin{aligned}
&P\left(\exists k' \in S^c,\ \left\|\mathbf{X}_{[k']}^T\tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n\right)\\
=&\mathbb{E}\left[P\left(\exists k' \in S^c,\ \left\|\mathbf{X}_{[k']}^T\tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n \middle| \tilde{\mathbf{R}}\right)\right]\\
\geq&\mathbb{E}\left[P\left(\exists k' \in S^c,\ \left\|\mathbf{X}_{[k']}^T\tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n \middle| \tilde{\mathbf{R}}\right)\mathbf{1}\left\{\sigma_{\min}(\tilde{\mathbf{R}}^T\tilde{\mathbf{R}}/n) > c\right\}\right]
\end{aligned}$$

Let $\mathbf{X}_{[k']}^{(d)}$ be the d-th column of $\mathbf{X}_{[k']}$. Then

$$\left\|\mathbf{X}_{[k']}^T\tilde{\mathbf{R}}\right\|_{\mathrm{F}}^2 = \sum_{d=1}^{D}\left\|\tilde{\mathbf{R}}^T\mathbf{X}_{[k']}^{(d)}\right\|_2^2 .$$

By our assumption, the entries of $\mathbf{X}_{S^c}$ are i.i.d. standard normal and independent of $\tilde{\mathbf{R}}$. Thus, by conditioning,

$$\tilde{\mathbf{R}}^T \mathbf{X}_{[k']}^{(d)} \Big| \tilde{\mathbf{R}} \sim \mathbf{Norm}(\mathbf{0}_D, \tilde{\mathbf{R}}^T \tilde{\mathbf{R}}) \ .$$

If $\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c$, meaning that $\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n \succeq c\mathbf{I}_D$, then $\left\|\tilde{\mathbf{R}}^T \mathbf{X}_{[k']}^{(d)}\right\|_2^2$ are stochastically larger than $cn\chi_D^2$, and independent across d's. Thus, $\left\|\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}\right\|_{\mathrm{F}}^2$ is stochastically larger than $cn\chi_{D^2}^2$, and hence

$$\begin{aligned}
&P\left(\left\|\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n \Big| \tilde{\mathbf{R}}\right) \mathbf{1}\left\{\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c\right\} \\
\geq & P\left(\chi_{D^2}^2 > n\lambda_n^2/c\right) \mathbf{1}\left\{\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c\right\} \ .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[P\left(\exists k' \in S^c,\ \left\|\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n \Big| \tilde{\mathbf{R}}\right) \mathbf{1}\left\{\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c\right\}\right] \\
\geq & \mathbb{E}\left[\left(1 - \prod_{k' \in S^c} P\left(\left\|\mathbf{X}_{[k']}^T \tilde{\mathbf{R}}/n\right\|_{\mathrm{F}} > \lambda_n \Big| \tilde{\mathbf{R}}\right)\right) \mathbf{1}\left\{\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c\right\}\right] \\
\geq & \left(1 - P\left(\chi_{D^2}^2 \leq n\lambda_n^2/c\right)^{p-s}\right) P\left(\sigma_{\min}(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}}/n) > c\right) \ .
\end{aligned}$$

We can bound the first term using **Lemma IV.9**. If $\log(p-s) > (z_0\alpha_1)^2$, and $\lambda_n < \frac{\sqrt{c}(D+\sqrt{\alpha_1^{-1}\log(p-s)})}{\sqrt{n}}$, then $1 - P\left(\chi_{D^2}^2 \leq n\lambda_n^2/c\right)^{p-s} \geq 1 - (1 - \frac{\alpha_2}{p-s})^{p-s} \geq 1 - e^{-\alpha_2} > 0$. Here z_0, α_1, α_2 are absolute constants defined in **Lemma IV.9**.

We then bound the second term using **Lemma IV.10**. Let $P_{[S]}^{\perp}$ be the orthogonal projection matrix onto the columns of $\mathbf{X}_{[S]}$, i.e.

$$P_{[S]}^{\perp} = \mathbf{I}_n - \mathbf{X}_{[S]}(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]})^{-1} \mathbf{X}_{[S]}^T \ .$$

Then we have:

$$\tilde{\mathbf{R}}^T \tilde{\mathbf{R}} \succeq \mathbf{Y}^T P_{[S]}^{\perp} \mathbf{Y} = \mathbf{E}^T P_{[S]}^{\perp} \mathbf{E} \ .$$

With probability one, $P_{[S]}^{\perp}$ has rank $n - sD$. Thus,

$$\mathbf{E}^T P_{[S]}^{\perp} \mathbf{E} \overset{d}{=} \sigma^2 \tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}} \ ,$$

where $\tilde{\mathbf{Z}}$ is a $n-sD$ by D Gaussian ensemble, each element follows $\mathbf{Norm}(0,1)$ i.i.d.. Thus

$$P(\sigma_{\min}(\tilde{\mathbf{R}}^T\tilde{\mathbf{R}}/n) > c) \geq P\left(\sigma_{\min}\left(\frac{\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}}{n-sD}\right) > \frac{cn}{\sigma^2(n-sD)}\right) .$$

We have assumed that $n - sD > (1+c_1)D$, and $n/sD > 1/(1+c_1)$. Let $c = \frac{c_1\sigma^2}{2(1+c_1)}\left(1-\sqrt{\frac{1}{1+c_1}}\right)^2$, then by **Lemma IV.10**,

$$P\left(\sigma_{\min}\left(\frac{\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}}{n-sD}\right) > \frac{cn}{\sigma^2(n-sD)}\right) \geq P\left(\sigma_{\min}\left(\frac{\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}}{n-sD}\right) \geq \frac{1}{2}\left(1-\sqrt{\frac{1}{1+c_1}}\right)^2\right)$$

$$\geq \alpha_4 .$$

In summary, choose $C_3 = \sqrt{\frac{c_1\sigma^2}{2(1+c_1)}}\left(1-\sqrt{\frac{1}{1+c_1}}\right)$, $C_4 = 1/\alpha_1$, and $C_5 = \alpha_4(1-\exp(-\alpha_2))$, the theorem has been proved. □

Here we only prove the lower bound for the case where the predictors and the noise are both i.i.d. This is the analytically most tractable case, for which we can provide a non-asymptotic bound: if the problem-related constants c_1, σ^2 are fixed, all the constants in the theorem can be specified, and we do not make vague requirement like "(n,p,s,D) is large enough". It is enough to prove a special case since we are dealing with lower bound here. The lower bound will hold as long as this simple case is included in the problem universe.

4.2.4 Compare to Lasso and group Lasso

In the literature, a typical choice of λ_n for classic Lasso regression is $\lambda_n \asymp \frac{\sqrt{\log p}}{\sqrt{n}}$; our choice is $\lambda_n \asymp \frac{D+\sqrt{\log p}}{\sqrt{n}}$. Asymptotically, this choice is the same as Lasso when $D = 1$ or $D = O(1)$.

This choice of λ_n also resembles the choice for group Lasso in the literature. Karim Lounici and Tsybakov (2011) use $\lambda_n \asymp \frac{\sqrt{D}+\sqrt{\log p}}{\sqrt{n}}$ for a group Lasso problem with p groups, each of size D. We group D^2 parameters in one block, so we have

D replacing $\sqrt{D}$. In the proof, we can see that λ_n is a high probability concentration bound, and it turns out that the dimension D is additive in the bound, not multiplicative.

As for the $\beta_{\min}$ condition, for Lasso, Wainwright (2009a) proved that it is sufficient to have $\beta_{\min} \asymp \lambda_n \vee \sqrt{\log s/n}$, as long as $n/s \log(p-s) > C_1'(1 + C_2'/s\lambda_n^2)$; if $\lambda_n \asymp \sqrt{\log p/n}$, then it is sufficient to have $\beta_{\min} \asymp \lambda_n$ as long as $n \succeq s \log p$. For our block-penalized regression, we have proved that it is enough to have $\beta_{\min} \asymp \lambda_n$, as long as $n \succeq s^2 D \vee s \log p$. Fixing D, our $\beta_{\min}$ condition matches that for Lasso asymptotically, but the sample size requirement $s^2 + s \log p$ can exceed $s \log p$ if $s \succ \log p$. However, we conjecture that the s^2 term is artificial and can be reduced to s. If the conjecture is correct, then we have a fully non-asymptotic theory that achieve the same rate as Lasso when fixing D, but can also deal with diverging D.

4.2.5 Disregard block structure

Even if the sparsity is block-wise, one can still neglect the block structure, and simply run D traditional Lasso regression to get element-wise sparse estimator. If all these Lasso regression achieves exact sparsity recovery element-wise, then overall the block-wise sparsity is also recovered exactly.

The comparison between these two approaches depends on the sparse structure of B. For example, we have the following three scenarios

1. $B = \frac{\beta_{\min}}{D} \mathbf{1}_{sD \times D}$, so that each block is a matrix where all entries are uniformly $\beta_{\min}/D$.

   If we disregard the block structure, we then have D regressions, with $s' = sD$, $p' = pD$ and $\beta'_{\min} = \beta_{\min}/D$. For one of the regression to have no false inclusion, we need $n' \asymp s' \log p'$ and $\lambda'_n \asymp \sqrt{\log(p'/n)}$; for all regressions to have no false inclusion, we need $n' \asymp s' \log(p'D)$ and $\lambda'_n \asymp \sqrt{\log(p'D)/n}$. This is because $\log p'$ come from union of p' events, and with D regressions, we have $p' \times D$ events. Similarly, in order to achieve no false exclusion, for one regression it is sufficient to have $n' \succeq s' \log(s')$, and then $\beta'_{\min} \asymp \lambda'_n$; for all D regressions to have no false exclusion the, we need to change the sufficient condition to $n' \succeq s' \log(s'D)$ instead of $s' \log(s')$ and $\beta'_{\min} \asymp \lambda'_n$. In summary, a sufficient sample size is $n' \succeq sD \log(pD^2) \vee \frac{D^2 \log(pD^2)}{\beta^2_{\min}}$.

2. $B = \frac{\beta_{\min}}{\sqrt{D}}[\mathbf{I}_D, \mathbf{I}_D, ..., \mathbf{I}_D]^T$, so that each block is a diagonal matrix with diagonal elements being $\beta_{\min}/\sqrt{D}$.

   The corresponding sparse level of Lasso is $s' = s$; dimension is $p' = pD$; $\beta'_{\min} = \beta_{\min}/\sqrt{D}$. Similar as before, a sufficient sample size is $n' \succeq s \log(pD^2) \vee \frac{D \log(pD^2)}{\beta^2_{\min}}$.

3. $B = \frac{\beta_{\min}}{\sqrt{s}}[..., \mathbf{e}_{d_{1k}d_{2k}}, ...]^T$, where $\mathbf{e}_{d_1 d_2}$ is a matrix whose (d_1, d_2)-th element is 1, and other elements are 0.

   The performance of using D Lasso regression depends on the positions of the non-zero elements in B. Suppose that the positions are chosen randomly. Then a sufficient sample size is $n' \succeq s \log(pD^2)/D \vee \frac{\log(pD^2)}{\beta^2_{\min}}$.

If D is fixed, then all these rates are the same with the one we conjecture, that is $s(D + \log p) \vee \frac{D + \log p}{\beta^2_{\min}}$ (the one we actually proved has $s^2 D$ in it, which can be worse). However, when D diverges, the conclusions vary by cases. In the first case, block-wise penalty is more effective, as the sparsity actually present in blocks; in the second case, block-wise penalty is more effective in general, unless $sD \succ s \log(pD^2) + D \log(pD^2)/\beta^2_{\min}$, which can hold if D is very large; in the third case, it is better to ignore group structure, as the true model is indeed element-wise sparse, i.e., only a few entries (in this case only one entry) are non-zero in each non-zero block.

4.3 Numerical study

Our theory suggests $\lambda_n \asymp \frac{D+\sqrt{\log p}}{\sqrt{n}}$, and for the probability of exact recovery to converge to 1, we have proved that it is enough to have sample size $n \succeq s^2 D + s\log p + \frac{D^2+\log p}{\beta_{\min}^2}$, although our conjecture is we only need $n \succeq sD + s\log p + \frac{D^2+\log p}{\beta_{\min}^2}$. We check the rate by running the following simulation. Data are generated by the following model:

$$\mathbf{X} = (\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}, ..., \mathbf{X}_{n\cdot})^T, \quad \mathbf{X}_{i\cdot} \overset{i.i.d.}{\sim} \mathbf{Norm}(0_{pD}, \Sigma) \ ,$$

$$\mathbf{E} = (\mathbf{E}_{1\cdot}, \mathbf{E}_{2\cdot}, ..., \mathbf{E}_{n\cdot})^T, \quad \mathbf{E}_{i\cdot} \overset{i.i.d.}{\sim} \mathbf{Norm}(0_D, \sigma_E^2 I_D \ ,)$$

$$\mathbf{Y} = \mathbf{X}B + \mathbf{E} = \mathbf{X}_{[S]}B_{[S]} + \mathbf{E} \ ,$$

where $B \in \mathbb{R}^{pD\times D}$ is the coefficient matrix and $B_{S^c} = \mathbf{0}_{(p-s)D\times D}$.

We consider different generative models:

1. Independent covariates $\Sigma = \mathbf{I}_{pD}$; coefficients $B_{[k]} = \frac{1}{\sqrt{sD}}\mathbf{1}_{D\times D}$, for $k \in S$.

2. Independent covariates $\Sigma = \mathbf{I}_{pD}$; coefficients $B_{[k]} = \frac{1}{\sqrt{sD}}\mathbf{I}_D$, for $k \in S$.

3. Independent covariates $\Sigma = \mathbf{I}_{pD}$; random coefficients $B^*_{[k]} \in \mathbb{R}^{D\times D}$ having i.i.d. Gaussian entries and $B_{[k]} = \frac{1}{\sqrt{s}}\frac{B^*_{[k]}}{\left\|B^*_{[k]}\right\|_{\mathrm{F}}}$.

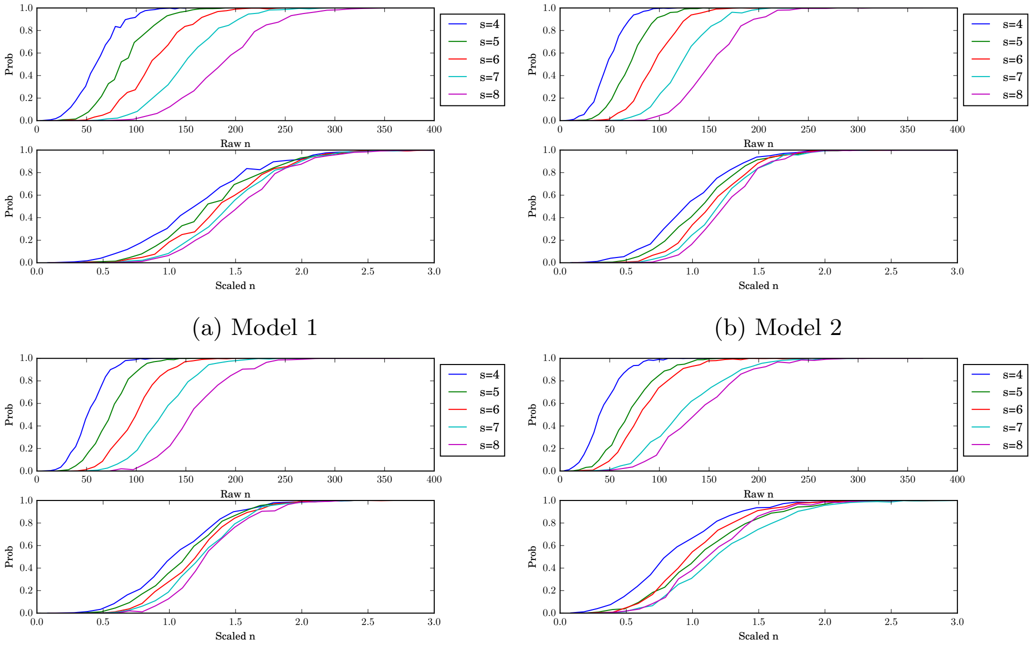
4. Dependent covariates $\Sigma = (\mathbf{I}_{p/4} \otimes M_{4\times 4}) \otimes \mathbf{I}_D$ where $M_{4\times 4} = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{pmatrix}$; coefficients $B_{[k]} = \frac{1}{\sqrt{sD}}\mathbf{1}_{D\times D}$, for $k \in S$.

Throughout this study, we fix $D = 3$. For each of the above models, we try different combinations of (s, p); S are s indices randomly selected from $\{1, 2, ..., p\}$. For each combination of (s, p), we try various sample sizes n and check the proportion of 500 independent runs where exact recovery $\hat{S} = S$ holds. We want to see how the empirical probabilities change with n. Penalty parameter is set to $\lambda_n = 0.5(D + \sqrt{3 \log p})/\sqrt{n}$.

For (s, p), we tested two sequences:

- Logarithmic sparsity: $(s, p) = (4, 32), (5, 64), (6, 128), (7, 256), (8, 512)$;
- Linear sparsity: $(s, p) = (2, 32), (4, 64), (8, 128), (16, 256), (32, 512)$.

The results are summarized in the Figure 4.1 and Figure 4.2. Each generative model and sparsity type has one plot corresponding to it and each plot has two panels: the upper panel shows how the probabilities change against raw sample size n, while the lower panel shows the probabilities against scaled sample size $n' = n/(sD^2 + s \log p)$.



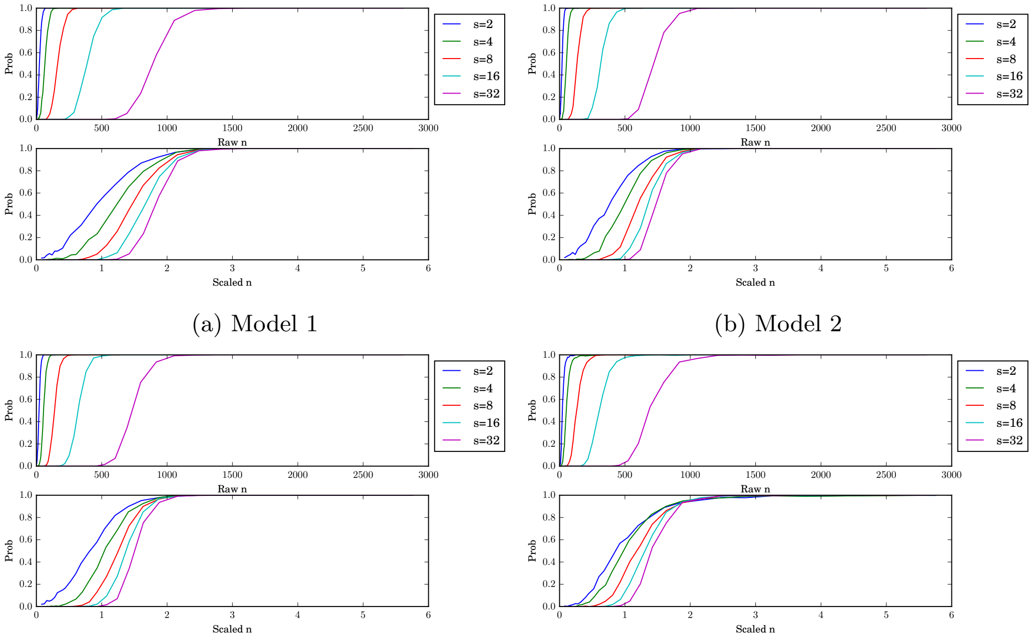
(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 4.1: Empirical probability of exact recovery against raw sample size vs. scaled sample size, logarithmic sparsity



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 4.2: Empirical probability of exact recovery against raw sample size vs. scaled sample size, linear sparsity

As s and p increase, the sample size n needed for the probability to get close to 1 also increases; however under all circumstances, the curves converge to 1 almost at the same scaled sample size n'. This validates our conjecture that when $\beta_{\min} \asymp 1/\sqrt{s}$, sample size needed for exact recover is of rate $s(D^2 + \log p)$, not $sD^2 + s^2D + s\log p$.

Note that in the experiments, the tuning parameter $\lambda_n = 0.5(D + \sqrt{3\log p})/\sqrt{n}$ is different from the theory. The constant multiplier 0.5 is picked arbitrarily rather than using $(1 + C_1)\sigma_E/(\gamma/2)$ (when $\Sigma = \mathbf{I}_{pD}$, it is something larger but close to 1), and we replace $2(\log(p - s) - \log\delta)$ with $3\log p$. The asymptotic conclusion is not affected.

4.4 Appendix

4.4.1 Auxiliary lemmas

We first derive a concentration bound regarding the cross product of two independent random matrices. This is used when bounding the cross product of the predictors and the noises.

Lemma IV.8 (Bounding the Frobenius norm of cross product of two random matrices)**.**

Suppose that we have two matrices $\mathbf{X} \in \mathbb{R}^{n\times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{n\times D_y}$. Assume $\frac{D_x \wedge D_y}{n} \le C$, for some constant $0 < C < 1$. Denote $W = \mathbf{X}^T\mathbf{Y}/n$.

1. *If $\mathbf{X}$ is a fixed matrix, and let $\Sigma_X^n = \mathbf{X}^T\mathbf{X}/n$. Moreover, $\mathbf{Y}$ is a random matrix, such that rows of $\mathbf{Y}$ are i.i.d. Gaussian vectors, $Y_i \overset{i.i.d.}{\sim} \mathbf{Norm}(0, \Sigma_Y)$. Then with probability at least $1 - \exp(-t)$,*

$$\sqrt{n}\|W\|_F \le \sqrt{\|\Sigma_X^n\|_{\mathrm{op}}\|\Sigma_Y\|_{\mathrm{op}}}(\sqrt{D_xD_y} + \sqrt{2t}) \ , \tag{4.19}$$

$$n\|W\|_F \le \|\mathbf{X}\|_{\mathrm{F}}\sqrt{\|\Sigma_Y\|_{\mathrm{op}}}(\sqrt{D_y} + \sqrt{2t}) \ . \tag{4.20}$$

2. If $\mathbf{X}, \mathbf{Y}$ *are two random matrices, where rows of* $\mathbf{Y}$ *are i.i.d. Gaussian vectors* $Y_i \overset{i.i.d.}{\sim} \mathbf{Norm}(0, \Sigma_Y)$*. Then for any* $c > 0$*:*

$$P\left(\sqrt{n}\|W\|_F \leq c\sqrt{\|\Sigma_Y\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t})\right)$$
$$\geq 1 - P(\left\|\mathbf{X}^T\mathbf{X}/n\right\|_{\text{op}} > c^2) - \exp(-t) \ .$$

Moreover, if rows of $\mathbf{X}$ *are also i.i.d. Gaussian vectors,* $X_i \overset{i.i.d.}{\sim} \mathbf{Norm}(0, \Sigma_X)$*. Then with probability at least* $1 - 2\exp(-s/2) - \exp(-t)$*,*

$$\sqrt{n}\|W\|_F \leq \left(1 + \sqrt{\frac{D_x \wedge D_y}{n}} + \frac{s}{\sqrt{n}}\right)\sqrt{\|\Sigma_X\|_{\text{op}}\|\Sigma_Y\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t}) \ .$$

Proof. 1. First assume $\Sigma_Y = I_{D_y}$, so that $\mathbf{Y}$ has i.i.d. entries following standard Gaussian. We have

$$\|W\|_F^2 = \frac{1}{n^2}\sum_{j=1}^{D_y}\left\|\mathbf{X}^T\mathbf{Y}_{\cdot j}\right\|_2^2 \ .$$

Note that $\mathbf{X}^T\mathbf{Y}_{\cdot j}/\sqrt{n} \sim \mathbf{Norm}(0, \Sigma_X^n)$, and these random vectors are independent across j. Let the eigenvalues of Σ_X^n to be $v_{x1} \geq ... \geq v_{xD_x}$. Then $v_{x1} = \|\Sigma_X^n\|_{\text{op}}$, $\sum_d v_{xd} = \|\mathbf{X}\|_{\text{F}}^2/n$. If $Z_{jd} \overset{i.i.d.}{\sim} \mathbf{Norm}(0,1)$, we have

$$\left\|\mathbf{X}^T\mathbf{Y}_{\cdot j}\right\|_2^2 \overset{d}{=} \sum_{d=1}^{D_x} v_{xd} Z_{jd}^2 \ .$$

Thus,

$$\|W\|_F^2 \overset{d}{=} \frac{1}{n}\sum_{j=1}^{D_y}\sum_{d=1}^{D_x} v_{xd} Z_{jd}^2 \ .$$

This will give us a chi-square like quantity. Use the bound in Laurent and Massart (2000), we get

$$P\left(n\|W\|_F^2 - D_y\sum_d v_{xd} \geq 2\sqrt{D_y\sum_d v_{xd}^2 t} + 2\max_d v_{xd} t\right) \leq \exp(-t) \ .$$

Note the two following inequality

$$D_y\sum_d v_{xd} + 2\sqrt{D_y\sum_d v_{xd}^2 t} + 2\max_d v_{xd} t \leq (\sqrt{D_y D_x} + \sqrt{2t})^2 v_{x1} \ ,$$

$$D_y \sum_d v_{xd} + 2\sqrt{D_y \sum_d v_{xd}^2 t} + 2 \max_d v_{xd} t \leq (\sqrt{D_y} + \sqrt{2t})^2 \sum_d v_{xd} \ .$$

Thus the event can be simplified to

$$P\left(n \left\|W\right\|_F^2 \geq (\sqrt{D_y D_x} + \sqrt{2t})^2 v_{x1}\right) \leq \exp(-t) \ ,$$

$$P\left(n \left\|W\right\|_F^2 \geq (\sqrt{D_y} + \sqrt{2t})^2 \sum_d v_{xd}\right) \leq \exp(-t) \ .$$

Note that $v_{x1} = \left\|\Sigma_X^n\right\|_{\text{op}}$, and $\sum_d v_{xd} = n\left\|\mathbf{X}\right\|_{\text{F}}^2$. The proof is completed.

In general, if $\Sigma_Y \neq \mathbf{I}_{D_y}$, then $\mathbf{Y} \overset{d}{=} \mathbf{Z}\Sigma_Y^{1/2}$, where $\mathbf{Z}$ is independent Gaussian ensemble. Then $\left\|\mathbf{X}^T\mathbf{Y}\right\|_{\text{F}} = \left\|\mathbf{X}^T\mathbf{Z}\Sigma_Y^{1/2}\right\|_{\text{F}} \leq \left\|\Sigma_Y\right\|_{\text{op}}^{1/2} \left\|\mathbf{X}^T\mathbf{Z}\right\|_{\text{F}}$. Proof achieved by multiplying both sides with $\left\|\Sigma_Y\right\|_{\text{op}}^{1/2}$

2. Let $\Sigma_X^n = \mathbf{X}^T\mathbf{X}/n$. Since $\mathbf{X}$ and $\mathbf{Y}$ are independent, we have

$$\begin{aligned}
&P\left(\sqrt{n}\left\|W\right\|_{\text{F}} \geq c\sqrt{\left\|\Sigma_Y\right\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t})\right) \\
=&\mathbb{E}\left[P\left(\sqrt{n}\left\|W\right\|_{\text{F}} \geq c\sqrt{\left\|\Sigma_Y\right\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t})|\mathbf{X}\right)\right] \\
=&\mathbb{E}\left[P\left(\sqrt{n}\left\|W\right\|_{\text{F}} \geq c\sqrt{\left\|\Sigma_Y\right\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t})|\mathbf{X}\right)\mathbf{1}\{\left\|\Sigma_X^n\right\|_{\text{op}} \geq c^2\}\right] \\
+&\mathbb{E}\left[P\left(\sqrt{n}\left\|W\right\|_{\text{F}} \geq \sqrt{\left\|\Sigma_X^n\right\|_{\text{op}}\left\|\Sigma_Y\right\|_{\text{op}}}(\sqrt{D_x D_y} + \sqrt{2t})|\mathbf{X}\right)\mathbf{1}\{\left\|\Sigma_X^n\right\|_{\text{op}} < c^2\}\right] \\
\leq& P(\left\|\Sigma_X^n\right\|_{\text{op}} \geq c^2) + \exp(-t) \ .
\end{aligned}$$

The first term is bounded because probability is always smaller or equal to 1; the second term is bounded because when conditioning on $\mathbf{X}$, the distribution of $\mathbf{Y}$ does not change, so we can use the high probability bound achieved in the first part.

By Lemma II.14

$$P\left(\left\|\Sigma_X^n\right\|_{\text{op}} > (1 + \sqrt{D_x/n} + s/\sqrt{n})^2 \left\|\Sigma_X\right\|_{\text{op}}\right) \leq 2\exp(-s^2/2) \ .$$

Combine these probabilities, we get with at least probability $1-2\exp(-s^2/2)-\exp(-t)$ that

$$\sqrt{n}\|W\|_F \leq (1+\sqrt{D_x/n}+s/\sqrt{n})\sqrt{\|\Sigma_X\|_{\text{op}}\|\Sigma_Y\|_{\text{op}}}(\sqrt{D_xD_y}+\sqrt{2t}) \ .$$

If $D_x > D_y$, we can exchange $\mathbf{X}$ with $\mathbf{Y}$ and obtain the same bound with D_y and D_x exchanged. This explains the $D_x \wedge D_y$ term in the final bound.

□

Here are some extra lemmas that are used when providing lower bound results.

Lemma IV.9 (Lower bound on chi-square tail)**.** *Denote* $H(d,z) = \mathbb{P}(\chi^2_d \geq (\sqrt{d}+z)^2)$*, then exists absolute constants* α_1, α_2, z_0 *(e.g.* $\alpha_1 = 1.362$*,* $\alpha_2 = 0.392$*,* $z_0 = 2.1$*) such that, for any* $z > z_0$*,*

$$H(d,z) \geq \alpha_2 \exp\{-\alpha_1 z^2\} \ .$$

Proof. Define

$$H(d,z) = \mathbb{P}(\chi^2_d \geq (\sqrt{d}+z)^2) = P(\sqrt{\chi^2_d}-\sqrt{d} \geq z) \ .$$

Consider $Y_d = \sqrt{X}-\sqrt{d}$, where $X \sim \chi^2_d$. Then p.d.f. of Y is

$$f_d(y) = \frac{1}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}(y+\sqrt{d})^{d-1}\exp\left\{-\frac{(y+\sqrt{d})^2}{2}\right\} \ .$$

We compare it with the density of $\mathbf{Norm}(0,1/2)$

$$f(y) = \frac{1}{\sqrt{\pi}}e^{-y^2} \ .$$

This is because when $d \to \infty$, $\sqrt{d}(\chi^2_d/d-1) \to Norm(0,2)$, so by the δ-method, $Y_d = \sqrt{d}(\sqrt{\chi^2_d/d}-1) \xrightarrow{d} \mathbf{Norm}(0,1/2)$.

We claim that exists some constant z_0, such that as long as $z > z_0$, $f_d(z) > f(z)$ holds for all d. To prove this claim, consider

$$\begin{aligned} h_d(z) &= \log(f_d(z)) - \log(f(z)) \\ &= (d-1)\log(z+\sqrt{d}) + \frac{z^2}{2} - \sqrt{d}z - \frac{d}{2} - (\frac{d}{2}-1)\log 2 - \log\Gamma(\frac{d}{2}) \ , \\ &= (d-1)\log\left(1+\frac{z}{\sqrt{d}}\right) + \frac{z^2}{2} - \sqrt{d}z - \frac{1}{6d} - C_d \end{aligned}$$

where

$$C'_d = \log\Gamma(\frac{d}{2}) + \frac{d}{2} - \frac{d}{2}\log(\frac{d}{2}) + \frac{1}{2}\log(\frac{d/2}{2\pi}) \ ,$$

$$C_d = C'_d - \frac{1}{6d} \ .$$

Note that

$$\frac{\partial h_d(y)}{\partial z} = \frac{d-1}{z+\sqrt{d}} + z - \sqrt{d} = \frac{z^2-1}{z+\sqrt{d}} \ .$$

Thus, $h_d(z)$ is increasing in z when $z > 1$. Using the inequality that $\log(1+x) > 1/(x+0.5)$ for $x > 0$, we have

$$\begin{aligned} h_d(z) &\geq (d-1)\frac{1}{\frac{\sqrt{d}}{z}+\frac{1}{2}} + \frac{z^2}{2} - \sqrt{d}z - \frac{1}{6d} - C_d \\ &= \frac{z^3-4z}{2(2\sqrt{d}+z)} - \frac{1}{6d} - C_d \ . \end{aligned}$$

We can prove that $C_d < 0$ for any $d \geq 2$ (using the Taylor expansion of log-gamma function). Thus let $z_0 = 2.1$

$$h_d(2.1) \geq \frac{0.861}{2(2\sqrt{d}+2.1)} - \frac{1}{6d} \ .$$

It is easy to show that the r.h.s. is larger than 0 if $d \geq 2$. Thus $h_d(2.1) > 0$. By monotonicity of $h_d(z)$, we have for any $d \geq 2$ that $h_d(z) \geq h_d(2.1) > 0$, which proves the claim.

With this claim, we have established that as long as $d \geq 2$, $z \geq 2.1$

$$H(d,z) \geq 1 - \Phi(\sqrt{2}z) \ .$$

A famous inequality is

$$1-\Phi(x) \geq \frac{x}{1+x^2}\phi(x) ,$$

where Φ, ϕ are c.d.f. and p.d.f. of standard normal, respectively. Thus,

$$\frac{\partial \log \Phi(-x)}{\partial x} = \frac{\phi(x)}{\Phi(-x)} \leq x(1+1/x^2) .$$

Substituting in $x=\sqrt{2}z$, and using the fact $z \geq 2.1$, we can find C so that $1+1/x^2 \leq 2C$. As an example, let $C=0.625$. Then

$$\frac{\partial \log \Phi(-x)}{\partial x} \leq 2Cx = \frac{\partial C x^2}{\partial x} ,$$

and hence $Cx^2+\log\Phi(-x)$ is a increasing function, so $Cx^2+\log\Phi(-x) > C'$ where C' is $Cx^2+\log\Phi(-x)$ evaluated at $2.1\sqrt{2}$. Therefore, if $z>2.1$, then we have

$$1-\Phi(\sqrt{2}z) \geq \exp(C'-2Cz^2) \geq 0.369e^{-1.25z^2} .$$

If $d=1$, $H(d,z)=P(\chi_1^2 \geq (1+z)^2) = 1-\Phi(1+z)$. Use a similar proof, and set $C=0.625$, C' is $Cx^2+\log\Phi(-x)$ evaluated at $1+2.1$. If $z>2.1$, we get

$$1-\Phi(1+z) \geq \exp(C'-2Cz^2) \geq 0.392e^{-0.625(1+z)^2} \geq 0.392e^{-1.362z^2} .$$

Note that the constants C, C' can be changed, but they are connected. The smaller C is, the smaller C' is. □

Lemma IV.10 (Lower bound on the minimal eigenvalue of the Gram matrix)**.** *If* $Z \in \mathbb{R}^{N\times M}$ *such that* $Z_{ij} \overset{i.i.d.}{\sim} \mathbf{Norm}(0,1)$*, and suppose* $N > N_0$*,* $M/N \leq \rho$*,* $0<\rho<1$*. If* $(1-\sqrt{\rho})^2 N_0 > \log 2$*, then*

$$P(\sigma_{\min}(Z^T Z/m) > \alpha_3) > \alpha_4$$

for some positive constants α_3, α_4*.*

Proof. By **Lemma II.14**,

$$
\begin{aligned}
2\exp(-CN_0) &\geq P\left(\sigma_{\min}(Z^T Z/N) \leq \left(1 - \frac{\sqrt{M}}{\sqrt{N}} - \frac{\sqrt{CN_0}}{\sqrt{N}}\right)^2\right) \\
&\geq P\left(\sigma_{\min}(Z^T Z/N) \leq (1 - \sqrt{\rho} - \sqrt{C})^2\right) .
\end{aligned}
$$

In order to avoid meaningless bound, we need $\sqrt{\rho} + \sqrt{C} < 1$ and $2\exp(-CN_0) < 1$. Such C can be found given the condition that $(1 - \sqrt{\rho})^2 N_0 > \log 2$. Then we can let $\alpha_3 = (1 - \sqrt{\rho} - \sqrt{C})^2$ and $\alpha_4 = 1 - 2\exp(-CN_0)$. □

4.5 Numerical study on $\left\|(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} \check{B}_{[S]}\right\|_{\infty,F}$

For simplicity, assume $D_1 = ...D_p = D_y = D$. We require in (4.8) that there exists constant C_2, t so that $\sqrt{s^2 D/n} + t\sqrt{s} < C_2$. From the proof, this is used to bound $\left\|(\mathbf{X}_{[S]}^T \mathbf{X}_{[S]}/n)^{-1} \check{B}_{[S]}\right\|_{\infty,F}$ from above. As pointed out in the end of **Remark IV.5**, we believe that this requirement is not necessary, and $n \succeq sD$ is enough.

When $D = 1$, the tighter rate can be proved using Levy's lemma on spherical concentration Ledoux (2005). The technique is used in Wainwright (2009a) for Lasso. When $D = 1$, with high probability $\check{B}_{[S]}$ is the true sign of β_S when λ is in the neighbourhood of 0; thus when increasing λ from 0, $\check{B}_{[S]}$ stays constant until λ passes some threshold. When $D > 1$ however, $\check{B}_{[S]}$ is random and hard to characterize.

Although we do not have the proof, numerical results do show that $n \asymp sD$ is enough, and here is our experiment. We fix $D = 3$, and vary $s = 2, 4, 8, 16, 32$. Note that the quantity of interest only involves the regression restricted to S, so p is not relevant. For a specific s, we let $n = 3sD$. $\mathbf{X}$ is generated so that the rows i.i.d. follow $\mathbf{Norm}(\mathbf{0}_{sD}, \Sigma_{[SS]})$, $\mathbf{E}$ is generated so that the rows i.i.d. follow $\mathbf{Norm}(\mathbf{0}_D, 0.5^2\mathbf{I}_D)$. The linear coefficients $B = \mathbf{1}_{sD\times D}$, and response is calculated using linear model $\mathbf{Y} = \mathbf{X}B + \mathbf{E}$. For the covariance matrix $\Sigma_{[SS]}$, we tested (i) an independent design

$\Sigma_{[SS]} = \mathbf{I}_{sD}$, (ii) a dependent design $\Sigma_{[SS]} = M \otimes \mathbf{I}_D$, where M is a s by s matrix so that $M_{kk} = 1$ for $k = 1, ..., s$, $M_{2l-1,2l} = M_{2l,2l-1} = 0.5$ for $l = 1..., s/2$ and all other entries being 0. Once data are generated, we compute the block penalized Lasso estimator $\hat{B}_{[S]}$, normalize each sub-block of $\hat{B}$ to get $\check{B}_{[S]}$, and then calculate the quantity of interest $W = \left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\right\|_{\infty,F}$. For each covariance model and each $(s, n = 3Ds)$ configuration, we generate 5000 independent data sets. The empirical distributions of W are given in Figure 4.3.

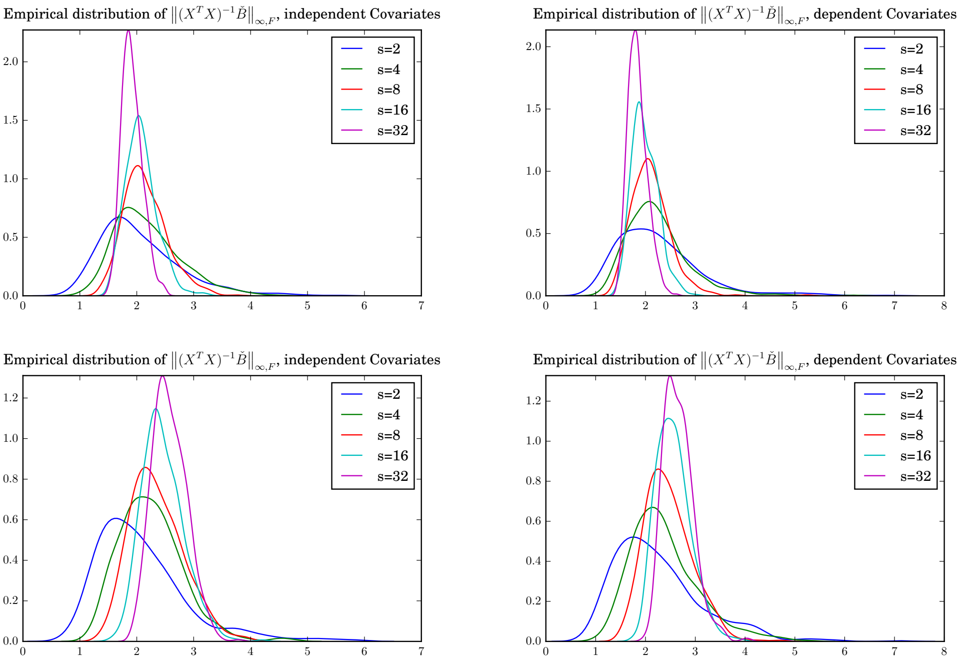


Figure 4.3: The distribution of $\left\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\right\|_{\infty,F}$ when s increases.

The first row corresponds to $\lambda_n = 0.1$ for all choices of s while the second row corresponds to $\lambda_n = 0.1/\sqrt{s/2}$, so that λ_n, as suggested by the theory, decreases as n increases. The plots on the left are results from identity covariance while on the right are those from non-identity covariance. Note that the values of λ_n are specified arbitrarily; the purpose of this experiment is to check whether the correlation between $(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]})^{-1}$ and $\check{B}_{[S]}$ causes W to be unbounded, so it is fine as long as λ_n is so large that $\hat{B}$ has too many zero blocks.

On the first row, clearly, the right tails of the distributions shrink rapidly as (s, n) increases; thus, if we want to check e.g. $P(\big\|(\mathbf{X}_{[S]}^T\mathbf{X}_{[S]}/n)^{-1}\check{B}_{[S]}\big\|_{\infty,F} < c)$ for $c = 3$, the probabilities actually shrink to 0. On the second row, although the modes shift right, the tails still shrink, so if we let $c = 4$, the probabilities still shrink to 0.

CHAPTER V

Future Research Directions

The current work has several limitations and leaves some open questions that need to be addressed in the future.

First of all, for thresholded SIR, theoretical results are provided only for the case where the rank $D = 1$ and the covariance of predictors $\Sigma_X = \mathbf{I}_p$. The results of thresholded PCA when $D > 1$ might be used to develop some analogous results for thresholded SIR when $D > 1$ and $\Sigma_X = \mathbf{I}_p$. For general Σ_X, SIR is related to a generalized eigenvalue problem which is more similar to canonical correlation analysis; in that case, some plug-in estimator of the precision matrix Σ_X^{-1} is inevitable. We can incorporate existing results on the property of the precision matrix estimator that is applied, but it is also interesting to discover an alternative approach to address the plug-in estimator more "organically".

Secondly, there are some practical issues that are not covered in this work. For example, a crucial aspect of dimension reduction is the estimation of rank D. Throughout this thesis, we assume D is known, which is rarely the case in reality. There are different criterions to decide the rank and it is interesting to compare their performances in SIR. Another possible future direction is to use some weighted norms instead of un-weighted ℓ-2 norm or Frobenius norm. In practice, it is often unclear how to scale the variables so that the coefficients in one group are equally important.

Finally, we have identified several limitations in the current proof techniques that result in some unsatisfactory terms in the error bounds or the sample size requirements. The difficulties in resolving these terms are usually the dependency between some random objects that are hard to characterize. This can lead to some deep theoretical researches which has not been accomplished.

Bibliography

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 4:1705–1732, 2009.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4.3:329–375, 1998.

A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41.3:1055, 2013.

E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse cca via precision adjusted iterative thresholding. *arXiv preprint*, arXiv:1311.6186, 2013.

X. Chen, C. Zou, and R. D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *the Annals of Statistics*, 38.6:3696–3723, 2010.

R. D. Cook. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91.435:983–992, 1996.

R. D. Cook. Save: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, 29.9-10:2109–2121, 2000.

A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, pages 41–48, 2005.

C. Gao, Z. Ma, and H. H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45.5:2074–2101, 2017.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38: 1978–2004, 2010.

B. Jiang and J. S. Liu. Variable selection for general index models vis sliced inverse regression. *The annals of Statistics*, 42.5:1751–1786, 2014.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104.486:682–693, 2009.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of computational and graphical statistics*, 12.3:531–547, 2003.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.

S. v. d. G. Karim Lounici, Massimiliano Pontil and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39: 2164–2204, 2011.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28.5:1356–1378, 2000.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

M. Ledoux. *The concentration of measure phenomenon.* American Mathematical Soc., 2005.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86.414:316–327, 1991.

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94.3:603–613, 2007.

L. Li and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48.4: 503–510, 2006.

L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrika*, 64: 124–131, 2008.

R.-C. Li. Relative perturbation theory: Ii. eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications*, 20.2:471–492, 1998.

Q. Lin, Z. Zhao, and J. S. Liu. On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv preprint*, arXiv:1507.03895, 2015.

Q. Lin, Z. Zhao, and J. S. Liu. Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, just-accepted, 2018.

H. Liu and J. Zhang. On the $\ell_1 - \ell_q$ regularized regression. *arXiv:0802.1517*, 2008.

Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41.2:772–801, 2013.

L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37:3779–3821, 2009.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *SIAM Journal on Numerical Analysis*, 34:1436–1462, 2006.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605C633, 2008.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27.4:538–557, 2012.

G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39:1–47, 2011.

H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99.6:1015–1034, 2008.

K. Tan, L. Shi, and Z. Yu. Sparse sir: optimal rates and adaptive estimation. 2017.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360C1392, 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

M. J. Wainwright. Thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on information theory*, 55:2183–2202, 2009a.

M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on information theory*, 55:5728–5741, 2009b.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10.3:515–534, 2009.

Y. Xia, H. Tong, W. K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64.3:363–410, 2002.

X. Yin and H. Hilafu. Sequential sufficient dimension reduction for large p small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77.4:879–892, 2015.

Z. Yu, L. Zhu, H. Peng, and L. Zhu. Dimension reduction and predictor selection in semiparametric models. *Biometrika*, 100.3:641–654, 2013.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468C3497, 2009.

W. Zhong, P. Zeng, P. Ma, J. S.Liu, and Y. Zhu. Rsir: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21.22:4169–4175, 2005.

S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *In Advances in Neural Information Processing Systems*, 22, 2009.

L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*, 106.496:1464–1475, 2011.

H. Zou and L. Xue. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15.2:265–286, 2006.

H. Zou, T. Hastie, and R. Tibshirani. A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106.8:1311–1320, 2018.